

Data Representation Scheme and Similarity Measures for a Comprehensive Computational Chemistry Database

by

© Mark Sinclair Staveley

A thesis submitted to the
School of Graduate Studies
in partial fulfilment of the
requirements for the degree of
Doctor of Philosophy

Department of Computer Science
Memorial University of Newfoundland

July 2009

St. John's

Newfoundland

Abstract

This thesis draws upon research in the areas of information retrieval, chemical informatics, and computational chemistry.

Many research initiatives deal with very large amounts of data, and as a result information retrieval systems are becoming more and more of a necessity. Chemically-based information retrieval systems are of particular interest to computational chemists, as computational chemists not only produce large quantities of information (data), but they also use large quantities of computer processing power (CPU cycles).

Currently there are no tools available through any of the Canadian High-Performance Computing consortia that have been designed and implemented to support the data management activities of computational chemists. The only electronic resources that are publicly available contain information that has either been obtained experimentally or through patent and publication searches.

A system by the name of Chem-DRSM has been designed and implemented in order to support the structuring and browsing of computational chemistry data. It has been implemented using principles and methods associated with various chemically based data representation schemes and similarity measures. This thesis presents and discusses the design, implementation and evaluation of the Chem-DRSM system. An evaluation of the similarity measures found within the Chem-DRSM system was conducted using statistical information (precision and recall statistics), information from

the distribution of similarity scores with test structures, and information gathered from a human study that involved subjects with an expert level of knowledge in chemistry.

The Chem-DRSM contains three different similarity measures (namely the contextual cosine measure, standard cosine measure, and Tanimoto measure), which have all been adapted to make use of specialized chemical topological descriptors called Chemical Atom Topological Indices (CATI). The evaluation not only compares the performance of these metrics with each other, but also compares their performance with a version of the Tanimoto measure which uses chemical fingerprints (which is considered to be an industry standard).

Results of the statistical evaluation showed that the standard cosine measure had a higher average precision (with a lower standard deviation) than the other measures (including the Tanimoto with chemical fingerprints). During the evaluation of the distribution of similarity scores produced by the different similarity measures it was observed that the standard cosine measure assessed the similarity of chemical structures with the most granularity. The level of granularity associated with the standard cosine measure is attributed, in part, to its use of statistical weighting information about the various descriptors found within chemical structures. This is in contrast to the Tanimoto measure, with chemical fingerprints, which only looks at the presence and absence of properties when distinguishing chemical structures. Furthermore, the standard cosine measure also identified more similar structures (as classified by the human study participants) than the Tanimoto measure.

All of these different evaluation results show that the standard cosine measure, which uses the CATI descriptors, defines a chemical information context for searching and browsing that is more appropriate than the chemical information context created by the Tanimoto measure which uses chemical fingerprints.

Dedication

This thesis is dedicated to the loving memory of my grandfather
Maj.(ret) Robert Malcolm “Mac” Sinclair, C’D, BSc(Mil), BComm. (1921-2006)
who always encouraged me, in the most meaningful and genuine ways, to put my
best foot forward and to have confidence in my own abilities.

Acknowledgements

Many people have taken the time to help me out while I have been working on this thesis, and I would like to draw attention to some of them. First, there is my wife Marcia Coleman. I am very grateful to Marcy for her support and understanding throughout this entire process as it is not easy having a husband that is going back to school. Second, there is our daughter Brigid. Brigid was very understanding about having to cut short playtime and also understanding when it was time for Daddy to work, thank you Brigid.

To my supervisors Dr. Ray Poirier and Dr. Sharene Bungay, I am especially grateful for the time they have spent assisting with the development and design of the Chem-DRSM system and I am also thankful for the way in which they encouraged me to not settle and to keep my standards high.

I would also like to say thank you to Debbie Earles, Bernice Devereaux, Elaine Boone, Michelle Shaw, Joshua Hollett, Phil Romkey, Sebastian Padina, Tim Richardson, John Warren, Frank Penney, Steve St.Clair, and Niven Sinclair for their discussions, assistance, and friendship.

Contents

Abstract	ii
Dedication	v
Acknowledgements	vi
List of Tables	xii
List of Figures	xvi
1 Introduction	1
1.1 Electronic Chemical Structure Archives	2
1.2 Thesis Outline	5
2 Chemical Data Representation	6
2.1 Introduction	6
2.2 Representing Chemical Information	7
2.2.1 Chemical Properties and Metadata Approach	8
2.2.2 Molecular Graph Theory Approach	10
2.2.3 Natural-Language / Chemical Semantics Approach	13

2.2.4	Quantum Chemistry-Based Approach	20
2.3	Summary	20
3	Similarity / Searching Methods and Techniques	22
3.1	Introduction	22
3.2	Properties and Metadata Search Methods	23
3.3	Similarity and Distance Coefficients	25
3.4	Fragment-Based Similarity Searching	29
3.5	Summary	32
4	The Chem-DRSM System	33
4.1	Multi-Component Data Representation Scheme (MC'DRS)	34
4.1.1	Chemical Atom Topological Index (CATI)	37
4.1.2	Chemical Bond Topological Indices (CBTI)	39
4.1.3	Nuclear Repulsion	44
4.1.4	Origin-Invariant Nuclear second-moment	45
4.1.5	Single Point HF/STO-3G Energy	46
4.1.6	InChI and SMILES descriptors	46
4.1.7	Metadata / Additional Properties	47
4.2	Implementation and design of the Chem-DRSM system	48
4.2.1	Structure Pre-Processing	51
4.2.2	Information Extraction	53
4.2.3	Index Creation	55
4.2.4	Chemically Based Similarity Measures	57
4.2.4.1	Vector Space Models	59

4.2.4.2	Similarity refinement using computationally derived chemical descriptors	64
5	Investigative Approach and Results	68
5.1	Determination of Similarity Threshold for Computationally Derived Descriptors	69
5.2	Statistical Evaluation	72
5.2.1	Choosing the test collection	73
5.2.2	Information extraction and index creation	74
5.2.3	Chemical structure similarity computations	74
5.2.4	Precision and Recall Evaluation	75
5.2.5	Data Analysis	76
5.3	Human Evaluation	91
5.3.1	Building on the Statistical Evaluation	92
5.3.2	Hypothesis	94
5.3.3	Subjects	94
5.3.4	Method	95
5.3.5	Data Collection	96
5.3.6	Data Analysis and Interpretation	102
5.4	Functional Group Investigation	106
5.5	Discussion	116
6	Prototype Comprehensive Computational Chemistry Database	118
6.1	Data Representation	119
6.2	Integration	119

6.3	Enhanced Chemical Information Classification	122
6.3.1	Practical Example of Clustering Groups of Chemical Structures - Carboxylic Acids	125
6.3.1.1	SMILES to SDF to XYZ	126
6.3.1.2	XYZ file Preparation	127
6.3.1.3	Computational Chemistry Calculations	128
6.3.1.4	Index Creation	128
6.3.1.5	Similarity Scoring	129
6.3.1.6	Relationship Determination	129
6.3.1.7	Discussion	130
6.4	Summary	131
7	Discussion and Future Work	132
7.1	Conclusions Drawn from Experimental Results	132
7.1.1	Precision-Recall Statistical Evaluation of Performance	133
7.1.2	Analysis of the Distribution of Similarity Scores	134
7.1.3	Results and Observations of Performance from the Human Eval- uation	135
7.2	System Performance Differences	136
7.2.0.1	Storage	137
7.2.0.2	Index Creation	138
7.2.0.3	Similarity Measure Calculations	139
7.3	Future Work	141
7.3.1	Extending the Human Evaluation	141

7.3.2	Further Experiments involving the Chem-DRSM system	142
7.3.3	Future Development and Applications	144
7.3.3.1	Chem-DRSM Version 2.0	145
7.3.3.2	National Comprehensive Computational Chemistry DataBase (NCCCDB)	146
	References	148
	A Statistical Data	157
	B Human Evaluation Data	177

List of Tables

4.1	Multi-Component Data Representation Scheme (MC'DRS) design goals.	35
4.2	Bond representation examples, as used in the Chemical Bond Topological Indices (CBTI).	41
4.3	Similarity engine design goals.	49
4.4	Information derived from the Cartesian coordinate representation of a chemical structure by the Chem-DSRM system.	54
5.1	Calculated thresholds for nuclear repulsion, origin-invariant nuclear second-moment, and STO-3G single point energy values.	71
5.2	Methods that were used to produce similarity scores.	75
5.3	Formulas, NSC numbers, number of structures that are identical, and the number of structures with the same formula in the test collection for the 19 different structures used in the statistical evaluation.	78
5.4	NSC numbers, precision values and statistical summary data for the 19 structures that were part of the statistical evaluation.	81
5.5	Distribution of similarity scores produced by different similarity measures when structure NSC 90799, $C_8H_{17}N$, is the query.	89

5.6	Distribution of similarity scores produced by different similarity measures when structure NSC' 167530, C_6H_6 , is the Query.	90
5.7	Gender breakdown of study participants.	94
5.8	Educational background of study participants.	95
5.9	Area of expertise of study participants.	95
5.10	Summary of similarity measure ordering for subject tasks.	97
5.11	Ordering of structures used to produce lists for each similarity measure that were assessed by study participants.	97
5.12	Number of similar structures that were identified by study participants.	104
5.13	The average user-assessed list correctness score, (and standard deviation), for each of the lists produced in the human evaluation.	105
5.14	Functional group listing and the information required for identification.	107
7.1	Differences in Storage Requirements for Different Indexing Schemes when Indexing 178,175 Structures.	137
A.1	Distribution of similarity scores produced by different similarity measures when structure NSC' 131564 is the query.	158
A.2	Distribution of similarity scores produced by different similarity measures when structure NSC' 134422 is the query.	159
A.3	Distribution of similarity scores produced by different similarity measures when structure NSC' 134438 is the query.	160
A.4	Distribution of similarity scores produced by different similarity measures when structure NSC' 152324 is the query.	161

A.5	Distribution of similarity scores produced by different similarity measures when structure NSC' 153096 is the query.	162
A.6	Distribution of similarity scores produced by different similarity measures when structure NSC' 167530 is the query.	163
A.7	Distribution of similarity scores produced by different similarity measures when structure NSC' 1765 is the query.	164
A.8	Distribution of similarity scores produced by different similarity measures when structure NSC' 169899 is the query.	165
A.9	Distribution of similarity scores produced by different similarity measures when structure NSC' 170347 is the query.	166
A.10	Distribution of similarity scores produced by different similarity measures when structure NSC' 209826 is the query.	167
A.11	Distribution of similarity scores produced by different similarity measures when structure NSC' 210746 is the query.	168
A.12	Distribution of similarity scores produced by different similarity measures when structure NSC' 15309 is the query.	169
A.13	Distribution of similarity scores produced by different similarity measures when structure NSC' 1880 is the query.	170
A.14	Distribution of similarity scores produced by different similarity measures when structure NSC' 525079 is the query.	171
A.15	Distribution of similarity scores produced by different similarity measures when structure NSC' 623441 is the query.	172
A.16	Distribution of similarity scores produced by different similarity measures when structure NSC' 26613 is the query.	173

A.17	Distribution of similarity scores produced by different similarity measures when structure NSC' 79367 is the query.	174
A.18	Distribution of similarity scores produced by different similarity measures when structure NSC' 8134 is the query.	175
A.19	Distribution of similarity scores produced by different similarity measures when structure NSC' 90799 is the query.	176
B.1	ID number, gender, education and area of specialization of human study participants.	178

List of Figures

2.1	Example MDL file for Sodium Fluoride (FNa)	10
2.2	Example of two different 3D structures being converted to the same 2D graph representation ($C_{10}H_{11}O_3$).	12
2.3	Four different SMILES representations for $C_3H_4O_2$	14
2.4	InChI representation for $C_3H_4O_2$	17
2.5	CML example for $C_3H_4O_2$	19
3.1	Partial fingerprint for $C_4H_{10}OS$	27
3.2	Partial fingerprint comparison of $C_4H_{10}OS$ and $C_6H_{14}O_2S$	28
3.3	Maximum common subgraph example.	31
4.1	Multi-Component Data Representation Scheme - Overview.	36
4.2	CATI for C_6H_6 example 1.	39
4.3	CATI for C_6H_6 example 2.	40
4.4	Chemical bond topological indices (CBTI) example differentiating be- tween a structure that has a carboxyl functional group (a) and a struc- ture that has an alcohol functional group (b).	43
4.5	Modular architecture of the Chem-DRSM system.	50

4.6	Example Cartesian coordinate and sdf representations for $C_3H_4O_2$, as well as a molecular graph of the structure for comparison.	52
4.7	Example of the indices created for the canonical SMILES descriptor. .	56
4.8	Architecture of the similarity measures found within the Chem-DRSM system.	57
4.9	Example of similarity measure scalability within the Chem-DRSM system.	58
4.10	CATI listing, with quantities, for $C_4H_{10}OS$ and $C_6H_{14}O_2S$	63
4.11	Example illustrating different nuclear repulsion, origin-invariant nuclear second-moment and single point energy values (RHF/STO-3G) for different $C_{22}H_{14}$ structures.	67
5.1	Calculation methodology showing the different optimized geometries that are used to determine the nuclear repulsion threshold, the origin-invariant nuclear second-moment threshold, and STO-3G single point energy threshold for any given structure (X).	70
5.2	Histograms of results (similarity scores) for query structure NSC 131564 with 4 similarity measures (as indicated)	83
5.3	Histograms of results (similarity scores) for query structure NSC 134422 with 4 similarity measures.	84
5.4	Histograms of results (similarity scores) for query structure NSC 134438 with 4 similarity measures.	85
5.5	Histograms of results (similarity scores) for query structure NSC 152321 with 4 similarity measures.	86

5.6	Query structures that were used to produce similarity scores for the human evaluation.	93
5.7	Sample pairwise similarity scoring web-form as used in the human evaluation.	99
5.8	Sample correctness of list scoring web-form as used in the human evaluation.	100
5.9	Sample e-mail generated from the web-based study interface.	101
5.10	CATI and CBTI descriptors used to identify the presence of amines, nitro compounds, and thiols within chemical structures.	109
5.11	CATI and CBTI descriptors used to identify the presence sulfides, nitriles, and aldehydes within chemical structures.	110
5.12	CATI and CBTI descriptors used to identify the presence of carboxyl groups, esters, and thioesters within chemical structures.	111
5.13	CATI and CBTI descriptors used to identify the presence of ethers and halides within chemical structures.	112
5.14	CATI and CBTI descriptors used to identify structures that are alkenes, alkynes or alkanes.	113
5.15	CATI and CBTI descriptors used to identify structures that are either aromatic (e.g. C_6H_6) or that contain alcohols.	114
5.16	CATI and CBTI descriptors used to identify the presence of ketones, acyl halides, amides, or acid anhydrides within a chemical structure.	115
6.1	Comprehensive Computational Chemistry Database architecture, including Chem-DRSM components.	121

6.2	Example Structure A - $C_6H_5N_2O_2Cl$	124
6.3	Example Structure B - $C_{10}H_{11}N_2O_3Cl$	125
B.1	Tabulated human evaluation data for test structures 1 and 2 using the contextual cosine measure that is part of the Chem-DRSM system. . .	179
B.2	Tabulated human evaluation data for test structures 3 and 4 using the contextual cosine measure that is part of the Chem-DRSM system. . .	180
B.3	Tabulated human evaluation data for test structure 5 using the contextual cosine measure that is part of the Chem-DRSM system. . . .	181
B.4	Tabulated human evaluation data for test structures 1 and 2 using the standard cosine measure that is part of the Chem-DRSM system. . . .	182
B.5	Tabulated human evaluation data for test structure 3 and 4 using the standard cosine measure that is part of the Chem-DRSM system. . . .	183
B.6	Tabulated human evaluation data for test structure 5 using the standard cosine measure that is part of the Chem-DRSM system.	184
B.7	Tabulated human evaluation data for test structures 1 and 2 using the Tanimoto measure with CATI descriptors that is part of the Chem-DRSM system.	185
B.8	Tabulated human evaluation data for test structures 3 and 4 using the Tanimoto measure with CATI descriptors that is part of the Chem-DRSM system.	186
B.9	Tabulated human evaluation data for test structure 5 using the Tanimoto measure with CATI descriptors that is part of the Chem-DRSM system.	187

B.10	Tabulated human evaluation data for test structures 1 and 2 using the Tanimoto measure with chemical fingerprints that is part of the OpenBabel system.	188
B.11	Tabulated human evaluation data for test structures 3 and 4 using the Tanimoto measure with chemical fingerprints that is part of the OpenBabel system.	189
B.12	Tabulated human evaluation data for test structure 5 using the Tanimoto measure with chemical fingerprints that is part of the OpenBabel system.	190
B.13	First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 1 from the human evaluation.	191
B.14	First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 2 from the human evaluation.	192
B.15	First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 3 from the human evaluation.	193
B.16	First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 4 from the human evaluation.	194
B.17	First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 5 from the human evaluation.	195

B.18	First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 1 from the human evaluation.	196
B.19	First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 2 from the human evaluation.	197
B.20	First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 3 from the human evaluation.	198
B.21	First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 4 from the human evaluation.	199
B.22	First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 5 from the human evaluation.	200

Chapter 1

Introduction

It is in a person's nature to want to learn about new things, and library-type resources are useful tools that can assist with the learning process. There are many different ways of obtaining information. In one instance this might involve searching for books or documents relating to a particular idea or subject. In another, it could involve browsing through information archives.

As information is gathered and collected, a need arises for this data to be stored and organized so that it can be easily accessed by users. Conventional libraries have developed techniques and schemes over hundreds of years to do this very thing. However, electronic or digital libraries are still very much in their infancy. Consequently, questions arise as to how one should go about structuring, organizing and searching within these environments. Additional questions relating to the types of tools and systems available for such an organizational task are also posed.

This thesis addresses the problem of managing and organizing large collections of chemical structure data. A new system called Chem-DRSM (Chemical structure Data Representation Similarity Measure) is presented. The Chem-DRSM system has been designed to provide assistance to chemists and chemical researchers within an electronic environment. This system distinguishes itself in two ways. First, it uses a customized data representation scheme that contains atom-based and bond-based topological descriptors, as well as computationally derived information. Second, it uses various similarity metrics that have been designed to maximize the use of the information contained within the chemical data representation scheme. These two features enhance the usability of the electronic information found within chemical structures, which in turn aids with the organization and searching of electronic chemical structure archives.

1.1 Electronic Chemical Structure Archives

One of the most influential developments in the way information is stored and accessed, is the development of what is known today as the Internet or the World Wide Web (WWW). Since the development of the Internet, in the early 1980's, information repositories that were once isolated are now connected to other such repositories and libraries. Some people have drawn similarities between the Internet and the beginning of a world-based encyclopedia [1].

In an attempt to better organize collections of chemical structures, electronic databases of chemical structures have been created. Examples of this include the Protein Data

Bank (PDB) [2], the National Cancer Institute (NCI) database [3], PubChem [4], and the ZINC database [5]. One of the main advantages of these types of resources is their accessibility as you do not have to physically be at the location of the journal articles or at the location where the structures may be stored to obtain information about a given structure. However, the maintainers of such collections are presented with a difficult task when trying to obtain the same levels of structure and organization that are commonplace within a conventional library.

One particular problem that is present with electronic chemical structure collections is that the metadata, which aids in the classification of the chemical structures, is not always readily available. It is possible for chemical structures to be accompanied by metadata; however in many cases the electronic versions of chemical structures do not have the same amount of metadata as that found in specialized chemical reference documents.

Text mining [6], automatic document summarization [7], and keyphrase extraction [8] are examples of techniques that are applied to English language text documents to solve this metadata problem. In terms of chemical structures, this problem is solved by using a number of different methods, including the translation of the chemical structure into some type of language-based representation (e.g. SMILES [9, 10] or InChI [11]), some type of molecular graph or topological based representation (e.g. Maximum Common Subgraph (MCS) [12]), or some type of representation that is based on computational chemistry (e.g. quantum chemistry) [13]. By combining the computed information from the chemical structure with any metadata that might be

available, chemical structures within an electronic collection can be better structured and organized.

By relying on methods that involve little or no human interaction to identify the information found with chemical structures, the structuring and organization of collections of chemical structures has become easier. Furthermore, by relying primarily on the topologically and computationally derived information as a basis for creating an organizational structure, which is representative of the chemical structures in the collection, the task of updating the electronic collection can be accomplished much faster as compared to the methods employed by traditional libraries. This is a good feature, as electronic collections tend to require updating on a more frequent basis because the contents of the collection are obtained from a variety of different sources (e.g. experimental, theoretical, computational, publications, etc).

Being able to browse and search effectively in an electronic environment can be a difficult task because the organizational structure and expertise found within a conventional library is not there. As chemical structures are converted into an electronic format, an opportunity to maximize information preservation and organization is presented to the curators of digital archives. For example, different descriptors can be obtained and calculated during the conversion process of a chemical structure into an electronic format. A review of some of these electronic descriptors and data representation schemes is presented in Chapter 2.

1.2 Thesis Outline

This thesis presents and investigates a method of electronically representing the information found within a chemical structure. Furthermore, it uses this information to determine chemical similarity using a number of different similarity measures. Both a precision-recall statistical evaluation and a human based study have been carried out to determine how well the information stored within the data representation scheme can be used to determine chemical similarity when using adaptations of standard information retrieval metrics.

Other than the Introduction, this thesis consists of six chapters (2-7). Chapter 2 reviews background work in the area of data representation and chemical structure information. Chapter 3 reviews and discusses similarity measures and similarity coefficients in terms of chemically-based information retrieval. Chapter 4 presents the Chem-DRSM system, which uses information from chemical structures to support information structuring, searching, and browsing activities. Chapter 5 discusses an investigation designed to evaluate the performance of the Chem-DRSM system, and presents the results that have been obtained. Chapter 6 draws from the observed results and the architecture of the Chem-DRSM system to present a design for a comprehensive computational chemistry database. Finally, Chapter 7 presents a discussion of the work within the thesis entire thesis as well as areas for future work.

Chapter 2

Chemical Data Representation

2.1 Introduction

Many different data representation schemes have been developed in order to process and store chemical information. Some of the more common data representation schemes have been created with a specific purpose or user group in mind. As an example, chemists, physicists and biochemists may use a searching and browsing based data representation scheme such as SMILES (Simplified Molecular Input Line Entry Specification) when they are searching large databases for structures, compounds, or scaffolds.

However, once results that are of interest have been found, the same users are then required to change data representation schemes in order to continue their work. One reason for changing data representation schemes might be the requirement to perform additional calculations on a given structure. In this case, a data representation scheme

that has preserved and contains the three-dimensional information of the structure (for example, a Cartesian coordinate file or Z-matrix) is required.

This need to change data representation schemes highlights one of the ways in which there is a lack of continuity between how chemical data is stored and how it is accessed. In some cases, chemical information repositories only store one or two of the different data representation schemes and scientists are therefore required to employ additional tools, such as conversion software like OpenBabel [14, 15], to complete their work.

This chapter presents a review of common data representation formats that store chemically based information, namely Chemical Properties and Metadata (2.2.1), Molecular Graph Theory (2.2.2), Natural-Language and Chemical Semantics (2.2.3), and Quantum Chemistry (2.2.4).

2.2 Representing Chemical Information

There are many reasons why chemical data would need to be searched. Examples include the need to search for a particular structure or sub-structure, wanting to screen structures for certain properties and characteristics, or perhaps trying to complete a patent search for a drug or related chemical structure. Whatever the reason for the search, the data representation scheme storing the chemical structure needs to be able to capture enough information about the chemical structure to allow these types of searches to be completed in a fast, efficient and reliable manner.

Common data representation schemes for chemical structures typically involve one or more of the following strategies: the use of metadata and other descriptors relating to the structure; the use of graph theory; the use of chemical nomenclature combined with English language rules and constructs; and the use of quantum chemistry-based descriptors (e.g. origin-invariant nuclear second moment and single point energy).

2.2.1 Chemical Properties and Metadata Approach

Metadata is not only limited to books or articles; metadata can be used to enhance the description of anything. However, it is possible to extract too much metadata and thereby capture redundant information. Chemical structures have a large number of properties and information that can be considered metadata. Molecular descriptors are good sources for chemical structure metadata as they are numerical values that characterize the properties of a molecule. Feher and Schmidt [16], as an example, use molecular properties to examine the differences between different compound classes (natural products, molecules from combinatorial synthesis, and drug molecules). Classification of chemical structures based on this kind of information is of particular interest to researchers in the area of combinatorial chemistry when trying to identify lead candidates for drug discovery. In the study conducted by Feher and Schmidt, 40 different properties were used when comparing structures. Examples of such properties include, the number of carbon-nitrogen bonds, the number of nitrogen atoms, the normalized number of ring systems, and the ring fusion degree.

There are a wide range of properties that can be derived from chemical structures. Some of the properties are more complex than others, and provide varying degrees of information that can be used to distinguish structures from one another. Some of the properties represent values that are simply the number of times a certain feature occurs within the structure. These are known as simple counts and examples include the number of carbon-carbon bonds, the number of carbon-nitrogen bonds, the number of rings, and the number of heavy atoms. However, simple counts are only the beginning when it comes to descriptors of chemical structures. Descriptors can be more advanced and subsequently require more information processing (time and computer power). Examples of advanced descriptors include topological indices [17-21], molar refractivity [22, 23], kappa shape indices [24] and electrotopological state indices [25].

Although most properties and descriptors are derived from the information contained within the chemical structure or the experimental conditions surrounding its creation, there are some descriptors and properties that have no relationship to the structure at all. These descriptors are sometimes referred to as dummy numbers and one of the most well known examples of a dummy number is the Chemical Abstracts Service (CAS) number that is associated with a particular structure [26]. Just as student numbers are assigned based on an enrollment or registration ordering, the CAS numbers are assigned sequentially as structures are registered in the Chemical Abstracts Service database.

2.2.2 Molecular Graph Theory Approach

Whereas molecular properties are considered to be derived from the chemical structure, the graph theory approach can be thought of as a means of translating the shape and connectivity of the chemical structure into an alternate representation.

Chemical structures can be stored in a number of different file formats and representations. These file formats typically contain information such as the three-dimensional coordinates of each of the atoms within the structure and information about how the atoms are connected. The MDL (Molecular Design Limited) [27, 28] and SDF (Structure Data File) [28] file formats are good examples of types of file format where both structural and connectivity information is represented. However, file formats with this information do not typically lend themselves to searching, as can be seen in Figure 2.1. In order to better support searching and browsing, the structures are translated into some type of a data structure that can support graphs and trees.

```
Test_NCI/62500/62526 NCI
OpenBabel05220913423D

  2  1  0  0  0  0  0  0  0  0999 V2000
  -1.0622  0.0000  0.0000 F  0  0  0  0  0
   0.8778  0.0000  0.0000 Na 0  0  0  0  0
  1  2  1  0  0  0
M END
```

Figure 2.1: Example MDL file for Sodium Fluoride (FNa)

Such a translation process treats the atoms within the structure as vertices and the bonds within the structure as edges. Once the chemical structure has been translated into a molecular graph representation, desired components and common structural fragments can be identified and searched for.

Although molecular graph theory [29, 30] lends itself quite well to chemical structures, there are a number of problems. The most challenging problem relates to how the graph should be ordered. Depending on the starting point, the same chemical structure can generate graph representations that have the nodes within the graph numbered differently. To solve this problem, a standard (or canonical) ordering scheme can be applied during the creation of the graph representation [31, 32]. As an example, Jochum and Gasteiger [31] use the following criteria to canonically number atoms of a molecular graph: atomic number, number of free electrons, number of atoms, atomic number of neighbours, number of bonds, bond priority, and the bond order.

Even with the canonical ordering of atoms within a molecular graph, there is still ambiguity in how the chemical structures are represented. Figure 2.2 shows an example of two chemical structures and their translation to a graph. As can be seen, the two different chemical conformations are represented by the same two-dimensional graph.

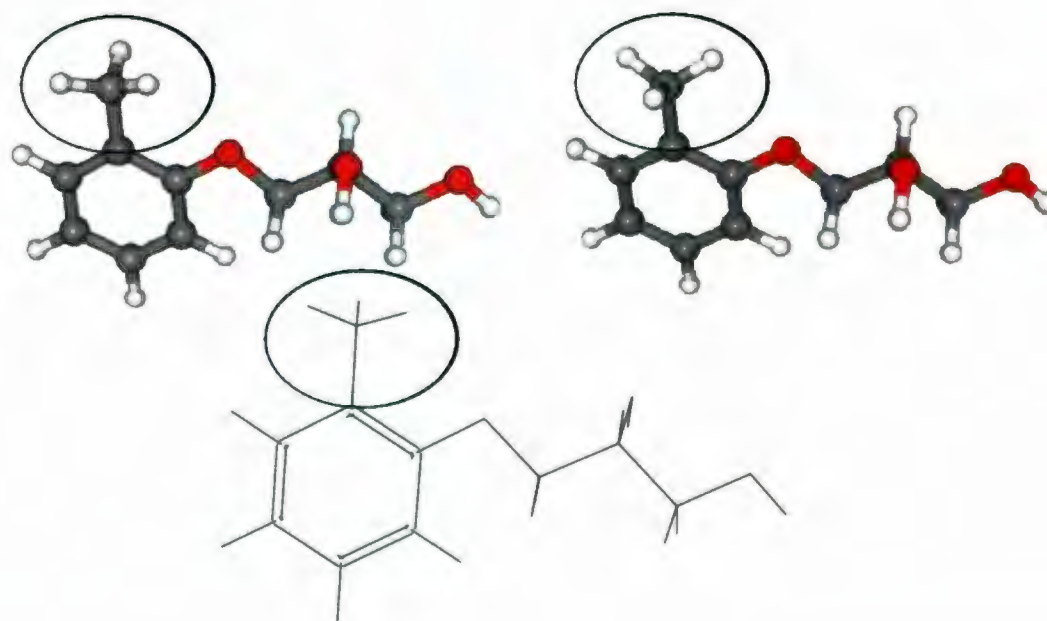


Figure 2.2: Example of two different 3D structures being converted to the same 2D graph representation (C₁₀H₁₄O₃). (Note the position of the atoms within the CH₃ group that is connected to the ring component of the structure).

2.2.3 Natural-Language / Chemical Semantics Approach

An alternative to the extraction of chemical properties and molecular graph theory is the combination of English language constructs and chemical semantics to represent chemical information. The most common representations to use this approach are the Simplified Molecular Input Line Entry Specification (SMILES) [9, 10], the Universal Chemical Key (UCK) [33], the IUPAC International Chemical Identifier (InChI) [11] and the Chemical Markup Language (CML) [34].

Typically, these methods involve a similar translation process to that used with molecular graphs. The structure is translated into its textual (or linear) representation as it is being traversed. In the case of SMILES, the translation process also involves the removal of Hydrogen atoms from the molecular graphs and further parsing of the molecular graph by a linguistic grammar. The SMILES grammar can deal with many different chemical semantics and various structural features. For example double bonds are represented by “=” and triple bonds are represented by “#”. Figure 2.3 shows examples of four different SMILES orderings that can represent $C_3H_4O_2$. The SMILES representation also has the ability to capture stereochemical relationships. For example, the configuration around double bonds can be specified by using “\” and “/”, and the configuration at a tetrahedral carbon is specified using either “@” or “@@” which represent counter-clockwise and clockwise traversals.

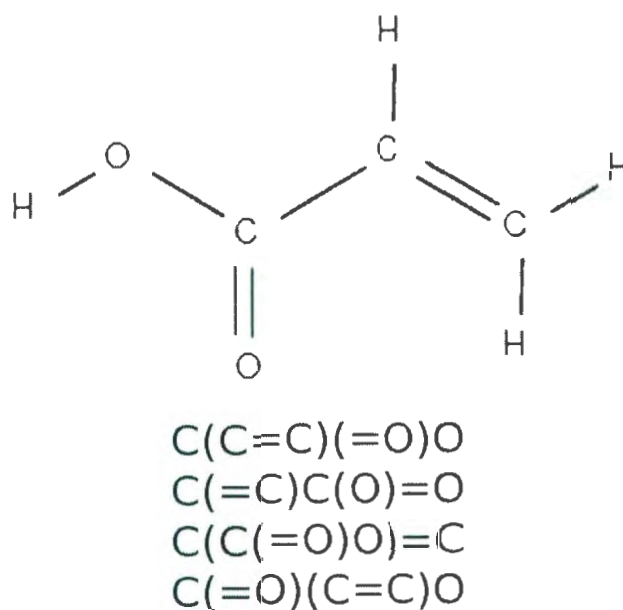


Figure 2.3: Four different SMILES representations for $C_3H_4O_2$

The SMILES system has been further extended by James. Weininger and Delany [35] to allow for the specification of patterns within chemical structures (as opposed to a single structure). This pattern specification language is referred to as SMILES Arbitrary Target Specification (SMARTS). SMARTS is a language, based on the SMILES natural-language grammar, that allows the user to specify substructures and patterns. For example, `[#6][CX3](=O)[($[OX2H0]([#6])[#6]),$([#7])]` is a SMARTS expression that matches an ester or amide [36].

Although these methods are able to capture a reasonable amount of the information associated with a chemical structure, it is possible to have an ambiguous representation. If the SMILES string is constructed using molecular graphs that are ordered differently, then it becomes possible to have different SMILES strings representing

the same chemical structure (Figure 2.3). To deal with this issue, a canonical ordering algorithm [10] is used with the SMILES construction process to produce what is referred to as "Unique SMILES" or "Canonical SMILES". This canonical ordering algorithm, called CANGEN, has two stages. The first stage of the algorithm, referred to as CANON, is used to label a molecular structure with canonical labels. In CANON the structure is treated as a graph with nodes (atoms) and edges (bonds), and each atom is given a numerical label on the basis of its topology. The second stage of the algorithm, referred to as GENES, generates a unique SMILES notation as a tree representation of the molecular graph. GENES selects the starting atom and makes branching decisions by referring to the canonical labels, as generated by the CANON algorithm. When the SMILES string has been canonically created, it is considered a unique SMILES string. The unique SMILES string for $C_3H_4O_2$ (as shown in Figure 2.3) is OC(=O)C=C.

A subsequent study by Grossman [33] has shown that the algorithm described in Weininger [10] does not create a unique SMILES string, and that it is possible to produce ambiguous results with relatively simple chemical structures. Example structures from Grossman's work include 1,3-diethyl-5-methylbenzene ($C_{11}H_{16}$ - NSC structure 62141). Grossman's work highlights how differences in the initial ordering of the atoms within the structure can lead to different unique SMILES strings when the CANGEN algorithm is used.

It is important to note however that the published CANGEN algorithm is not the same as the commercially available algorithm that the creators of the SMILES rep-

resentation (Daylight Chemical Information Systems Inc.) use for the generation of unique SMILES strings [37]. Consequently, due to lack of access to this commercial software, it is possible that the commercially based algorithm does not have the ambiguities that Grossman discovered. Although, since both SMILES representations are based on the use of a molecular graph, information is still lost when converting chemical structures to molecular graphs (three-dimensional to two-dimensional conversion).

In response to the ambiguities discovered Grossman went on to propose his own solution, called the Universal Chemical Key (UCK) [33]. The UCK starts by creating a labeled graph for the chemical structure where the labels contain information about the local connectivity of the structure. The next step is to combine the information contained within the labeled graph with information relating to the shortest path between every pair of atoms within the structure. This information is then concatenated to create a unique string, and for data storage purposes the string is further processed using the MD5 hashing algorithm. The two major disadvantages of this method are that the MD5 hashing algorithm can produce ambiguous results in that different inputs into the MD5 hashing algorithm can produce the same output [38], and the resulting MD5 string is not easily deciphered.

The International Union of Pure and Applied Chemistry (IUPAC) has also created their own chemical structure representation, called the International Chemical Identifier (InChI) [11]. In addition to using canonically ordered information about the structure (ordered in a way very similar to the CANGEN algorithm), this approach uses many different types of information about the chemical structure and stores this

in predefined layers. An example of an InChI string can be seen in Figure 2.4. As can be seen, the InChI representation is not easily parsed. Although the InChI representation was not designed to be interpreted by humans, some insight can be gained by looking at the different layers that make up the InChI representation as a whole. As pointed out within the InChI documentation [39], there are six different InChI layer types, each representing a different class of structural information. These layers are the main layer, the charge layer, the stereochemical layer, the isotopic layer, the fixed-H layer and the reconnected layer.

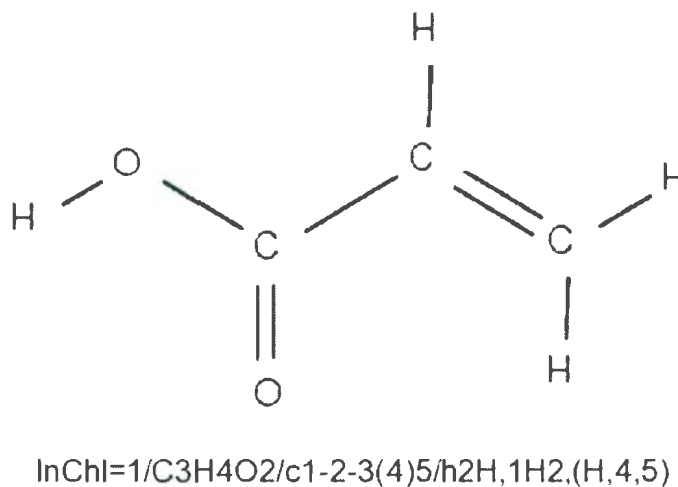


Figure 2.4: InChI representation for C₃H₄O₂

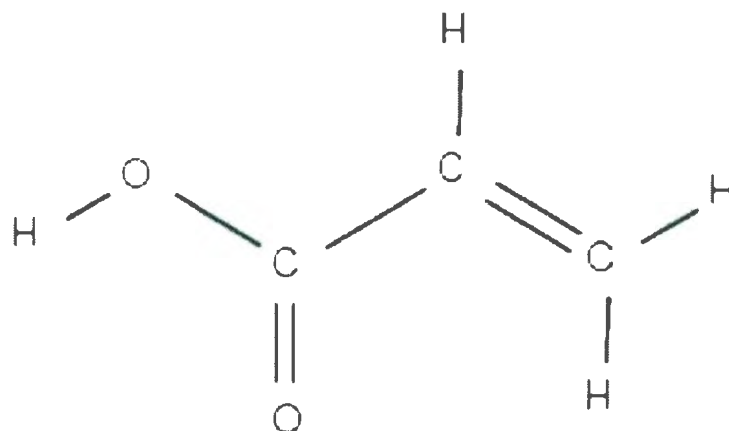
Different scientists have different needs when it comes to chemical structure information. InChI has been designed in such a way as to provide maximum flexibility in how it represents different information components. The flexibility of InChI can be

attributed, in part, to how the InChI identifier is represented. The identifier and associated text output may be parsed and annotated in either a simple plain text or XML.

The fact that the InChI identifier conforms to XML standards means that it can also become part of the structural representation generated using the Chemical Markup Language (CML) [31]. CML has been designed in such a way that it draws from the information representation techniques found within various markup languages (HTML, XML and SGML for example).

Markup languages have a fundamental concept, or building block, known as an identifier. When a person creates an HTML document they use various identifiers to structure, and then subsequently parse, the information within a document. CML attempts to provide a similar mechanism for the identification and structuring of information found within, and associated with, chemical structures. Figure 2.5 is an example of the CML representation of a chemical structure.

If one wanted to add InChI information to the CML example in Figure 2.5, then the following CML code could be added to the CML file: `<identifier convention="iupac inchi" value= "InChI=1/C3H4O2/c1-2-3(4)5/h2H,1H2,(H,4,5)">`. The ease with which the two can be combined demonstrates how the two information representation schemes can be adapted to compliment each other. The combination of InChI and CML has become a primary method for representing chemical information within environments that require XML standards.



```

<?xml version="1.0"?>
<molecule id="id4721_NCl.xyz" xmlns="http://www.xml-cml.org/schema">
  <name>4721_NCl.xyz</name>
  <atomArray>
    <atom id="a1" elementType="C" x3="0.854300" y3="-0.687000" z3="-0.000200"/>
    <atom id="a2" elementType="C" x3="-0.428200" y3="0.030400" z3="-0.000000"/>
    <atom id="a3" elementType="O" x3="-0.444700" y3="1.245700" z3="0.000200"/>
    <atom id="a4" elementType="O" x3="-1.586400" y3="-0.659300" z3="-0.000100"/>
    <atom id="a5" elementType="C" x3="1.997700" y3="-0.006000" z3="-0.000100"/>
    <atom id="a6" elementType="H" x3="0.868900" y3="-1.766900" z3="0.004100"/>
    <atom id="a7" elementType="H" x3="-2.430300" y3="-0.187200" z3="-0.000000"/>
    <atom id="a8" elementType="H" x3="2.940200" y3="-0.533300" z3="-0.000300"/>
    <atom id="a9" elementType="H" x3="1.983000" y3="1.073900" z3="0.000100"/>
  </atomArray>
  <bondArray>
    <bond atomRefs2="a8 a5" order="1"/>
    <bond atomRefs2="a1 a5" order="2"/>
    <bond atomRefs2="a1 a2" order="1"/>
    <bond atomRefs2="a1 a6" order="1"/>
    <bond atomRefs2="a4 a2" order="1"/>
    <bond atomRefs2="a4 a7" order="1"/>
    <bond atomRefs2="a5 a9" order="1"/>
    <bond atomRefs2="a2 a3" order="2"/>
  </bondArray>
</molecule>

```

Figure 2.5: CML example for $C_3H_4O_2$

2.2.4 Quantum Chemistry-Based Approach

All the methods reviewed thus far involve the identification and organization of information associated with chemical structures. The approach involving quantum chemistry is no different. Quantum chemistry uses the information contained within the wavefunction and density to better understand the properties and behaviour of a given chemical structure.

The main disadvantage of the quantum chemistry approach is that even the simplest of approximations using the wavefunction can be very costly in terms of CPU time and required computational resources. In contrast to this disadvantage, the only error associated with the quantum chemistry calculations is in the method itself. One particular advantage of this approach, is that based on the result from one calculation, many other molecular descriptors may subsequently be derived. This in turn can assist database designers as they decide what data values are to be stored and what data values are to be derived [40].

2.3 Summary

Upon the completion of the review of data representation schemes, it is concluded that although a great deal of chemical information is captured through the use of meta-data, chemical properties, molecular graphs, textual representations, and quantum descriptors, there are still deficiencies as chemical information can be lost and improperly approximated based upon the choices made for data representation. These deficiencies are highlighted by the changing needs of chemical researchers who are

becoming more and more dependent on specialized information management. For example, computational chemists need to be able to easily use the information that is contained within transition states, excited states, and conformers and this is information which is not easy (or in some cases, even possible) to represent with the data representation schemes described in this chapter.

Chapter 3

Similarity / Searching Methods and Techniques

3.1 Introduction

The quality of results obtained through searching and browsing activities is dependent, in part, on the method in which the data has been stored and organized. Over the years, librarians have played a vital role, not only in the organization and classification of information, but also in the area of data search strategies and search design. With the development of tools that organize and structure information (for example, the World Wide Web and database systems), information resources have been migrated to electronic warehouses and digital libraries. As a result, people are now able to access a wide range of information resources without being concerned about geographical proximity or restrictions.

Google [41], Bing [42] and the New Zealand Digital Library [43] are examples of the many different systems that are available to digitally store and access information. These three systems serve as example information repositories that are generic in nature as they have been designed to process multi-disciplinary information from multiple data sources (e.g. images, maps, text, books, etc). The Protein Data Bank [2] and the National Cancer Institute (NCI) Database [3], on the other hand, are examples of purpose built information management and retrieval systems that concentrate on a specific area or discipline. This chapter outlines a number of different searching methods designed to be used with the different chemical data representation schemes outlined in the previous chapter.

3.2 Properties and Metadata Search Methods

By using various search criteria, one can proceed to search electronic resources through the use of different query constructs. Boolean and ranked queries are the two most common types of queries in use.

Boolean queries use logical operators (such as AND, OR and NOT) to create logical search expressions. These expressions are then combined in various ways to return to the user a list of matching search results from the electronic data resource. Although Boolean query techniques are useful, they are unable to provide any ordering (or ranking) of the results. From the point of view of the retrieval system, all the search results are equally correct as they satisfy the Boolean expression contained in the query.

By using Boolean queries, it becomes possible to search for chemical structures using logical expressions that are based on metadata and other property information that has been extracted from the chemical structure. As an example, a search query could be constructed to find all of the structures that contain a carbon-carbon triple bond.

However, the use of a Boolean query puts the onus on the user to understand the properties being considered, and to create a suitable logical expression for searching. If a user wanted to use an automatic approach where the search was simply to find all the chemical structures that are similar to a given structure, then some type of a searching method that could provide a similarity or confidence rating would be required. If a Boolean search strategy was used, then the task of query refinement would be difficult as the searcher would have to reconstruct or add to their logical expression with each iteration.

In contrast to a Boolean query is the idea of a ranked query. Ranked queries search for results by looking at a set of properties and rather than trying to match an explicit logical expression, the information being searched is instead ranked based on how well the items best match the search criteria. The ordering is based on a confidence or similarity score that the search algorithm assigns to each item based on the query it is given.

3.3 Similarity and Distance Coefficients

The search specifications, which are contained within the query, need to be interpreted in a timely fashion so that the delivery of information can be done in a reasonable amount of time. As such, the algorithm used to generate the ranking score must not only be accurate, reliable, and consistent, but it must also be fast.

Information retrieval science has been investigating many different ranking mechanisms for use with English language text. One approach is to use information relating to the statistical occurrence of words within documents, as opposed to word meanings. Examples of this approach include the cosine coefficient [1], or the inner product coefficient [1]. Other methods, such as semantic indexing [14], attempt to better understand and use word meanings and their context. However, even when using advanced methods that utilize word semantics and meaning, there is still difficulty in capturing the author’s meaning and the linguistic context of the document.

Chemical properties, structure fragments, and metadata are considered to have less ambiguous information when compared with English language text. As a result, various mechanisms such as Wolfram Alpha [45] have been developed in an attempt to capture chemical information and subsequently provide accurate, fast, and reliable search tools based on this chemical information. When searching for chemical information, it is possible to use Boolean search techniques but this approach is limited in nature. Consequently, different ranking algorithms have been modified and adapted for use with various types of chemical information.

A review of different chemical similarity searching algorithms is presented by Willett, Barnard and Downs [46]. As part of the review, a performance comparison of similarity and distance coefficients by Willett and Winterman [47] was discussed. In this performance review, a metric known as the Tanimoto similarity measure [48] was deemed to be the preferred measure. The Tanimoto similarity coefficient is calculated using the following equation

$$Tanimoto = \frac{N_{a \cap b}}{(N_a + N_b - N_{a \cap b})} \quad (3.1)$$

where N_a represents the number of properties found in A, N_b represents the number of properties found in B and $N_{a \cap b}$ represents the number of properties common to both A and B. Although A and B can be any two things being compared, Willett and Wintermans evaluation of the Tanimoto similarity coefficient was conducted using chemical structures. The preference observed in the study was attributed partly to the bias of a subjective evaluation of the similarity measure, and partly because the calculation of the Tanimoto coefficient is not as complex as other measures, and as such is faster. When implementing the Tanimoto equation in chemical information management and retrieval systems, it is the user and the system designer that have the ability to determine what properties are considered when calculating similarity. This is in contrast to Boolean searches where the user is required to create the logical expression for searching.

A common approach for searching is to use a Tanimoto coefficient that has been calculated using chemical fingerprints. A fingerprint, in this context, is created by

combining a number of different properties in order to summarize the chemical structure. Some of the properties used within the chemical fingerprints include simple counts, and the presence of certain substructures. A list of all possible properties are treated as a vector with each property being assigned a particular bit. If that item is found within a given structure, then that particular bit is activated, otherwise the bit remains off. An example of how fingerprints are created can be seen in Figure 3.1, where various structural bits are activated for $C_4H_{10}OS$.

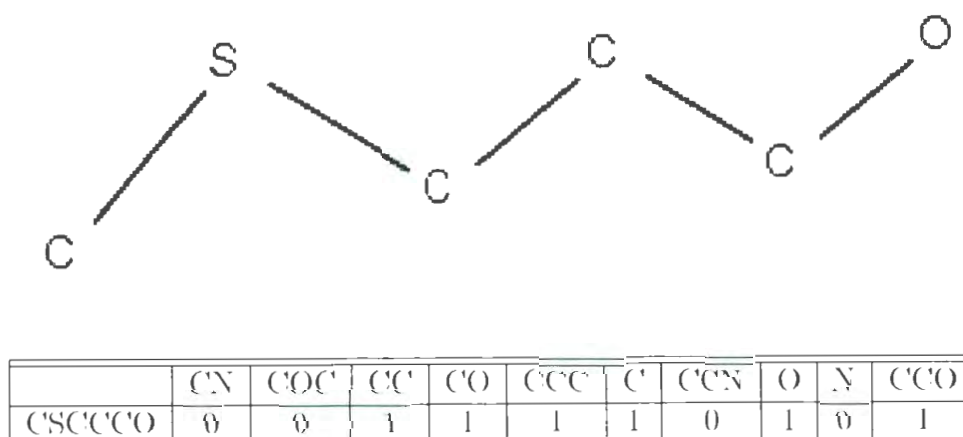
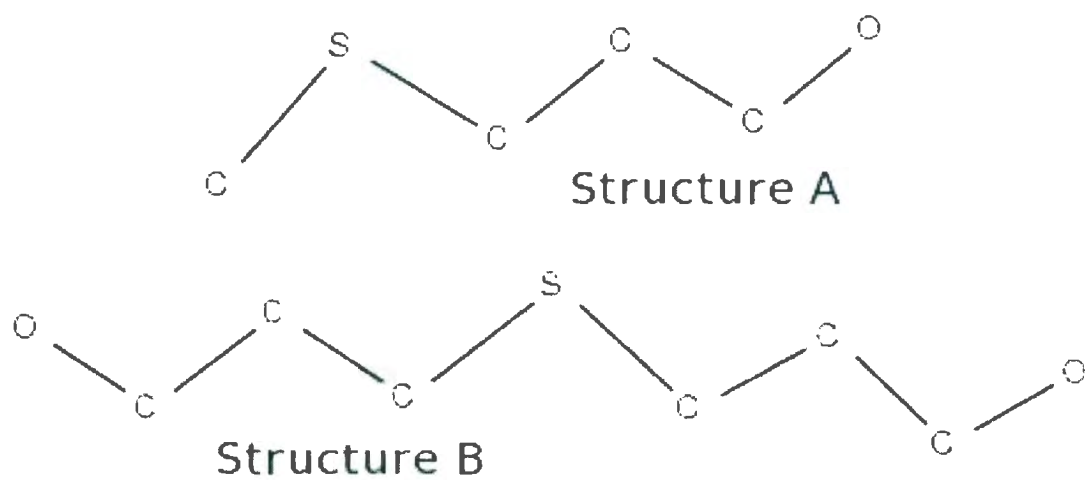


Figure 3.1: Partial fingerprint for $C_4H_{10}OS$. (Note: hydrogen atoms are not shown in structural representations).

Once the fingerprints have been determined for multiple structures, then the Tanimoto equation can be applied to assess chemical similarity. Figure 3.2 shows representative fingerprints for two different chemical structures, namely $C_4H_{10}OS$ and $C_6H_{14}O_2S$, as well as the intersection of the two fingerprints. Even though the Tanimoto equation produces a confidence score of 1.0 (100% similarity) when using the set of properties



	CN	COC	CC	CO	CCC	C	CCN	O	N	CCO
Structure A (C'SC'C'CO)	0	0	1	1	1	1	0	1	0	1
Structure B (OCC'C'SC'C'CO)	0	0	1	1	1	1	0	1	0	1
Structure A & B	0	0	1	1	1	1	0	1	0	1

Figure 3.2: Partial fingerprint comparison of $C_4H_{10}OS$ and $C_6H_{14}O_2S$. (Note: hydrogen atoms are not shown in structural representations).

in Figure 3.2, it is important to note that only the presence of properties within the chemical fingerprints, and not their quantities of occurrence, is represented (as can be seen with the (C'C'C') property). However, the use of the Tanimoto equation and chemical fingerprints is considered to be an industry standard in terms of assessing the similarity of chemical structures.

3.4 Fragment-Based Similarity Searching

The similarity measures outlined thus far do not provide any means to identify local regions of similarity between two structures [49]. As an example, a Tanimoto similarity score does not provide any insight into what common components two chemical structures might have. Rather, the similarity score is just a measure of confidence as to how similar the two structures are.

An alternative approach to calculating similarity scores, or coefficients, involves the generation of a mapping or alignment of the common components of two structures. An example of this approach is called the maximum common subgraph (MCS) [12] which is defined as the largest set of nodes (atoms) and edges (bonds) in common between two molecular graphs (structures). Figure 3.3 is an example of the type of information that is captured within a MCS, the MCS of structure A & B represents the structural components that are the same within the two different structures. It is important to note that the determination of the MCS for a given set of graphs is a NP-bound problem. As such, calculation times can increase drastically as the number of graphs being compared increases.

However, the information contained within the MCS can be used as a measure of similarity. For example the ratio between the size of the MCS and the size of the structures can provide insight into how similar two structures are. Additionally, the MCS can be used to assist with modeling reactions, as the MCS between products and reactants provides information as to where reaction activity (bonds being broken, bonds changing, atoms being removed, etc) is taking place [50].

Notable work in this area includes the development of a similarity algorithm by Raymond et. al. [51] in 2002 that uses various heuristics when identifying a MCS. One of the notable features of this algorithm is that it has the ability to perform tens of thousands of comparisons per minute because it uses specialized pre-screening techniques to reduce the number of MCS calculations that ultimately need to be performed.

Using the MCS as search criteria provides localized information about the chemical structures being considered, information such as receptor or docking sites which can subsequently provide more insight into the nature of the chemical structure. However, it is still limited by the inability of the graph representation to capture three-dimensional information (as can be seen in Figure 2.2, Section 2.2.2).

Even though the MCS does not capture all of the three-dimensional information relating to a chemical structure, it can be combined with other complimentary calculations for the purpose of determining chemical structure alignment which can be useful when determining chemical similarity. Work in the area of the automatic

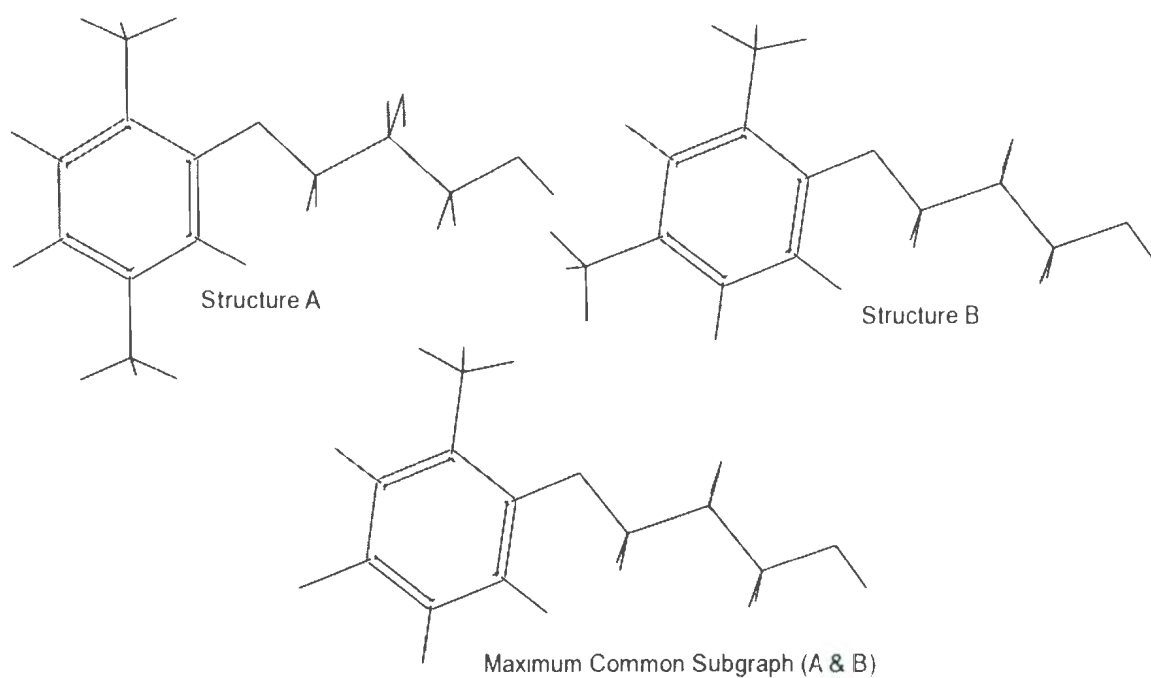


Figure 3.3: Maximum common subgraph example. (Note the only difference between structures A and B is the different ring substituents)

alignment of three-dimensional structures has been completed by Girones, Robert, and Carbo-Dorca [52], whose approach is based on the classification of atoms within the structure and interatomic distances. Duca and Hopfinger [13], on the other hand, have taken an alternate approach to this problem as they use the conformational energy profile of the molecule to assist with the determination of similarity.

3.5 Summary

Upon reviewing the different methods of assessing similarity and the associated searching techniques, it is concluded that a ranked measure would serve to offer the best balance between data representation scheme flexibility and computer processing requirements. The next step is to implement a system that makes use of a suitable data representation scheme and appropriate similarity measures. The next chapter in this thesis outlines the Chem-DRSM system which is a system that was created to support searching and browsing activities that relate to chemical structure data.

Chapter 4

The Chem-DRSM System

Computational chemists require detailed information about the three-dimensional nature of a chemical structure when performing calculations. One of the design challenges with this thesis has been to either create or find suitable methodologies that will allow for the efficient storage of three-dimensional information. The caveat is that this information must also be stored in such a way that it not only supports searching and browsing, but that it also provides users with results that are non-ambiguous. Furthermore, the information being returned must preserve all of the structure's information (e.g the three dimensional information of the structure and not a two dimensional molecular graph or projection).

SMILES, InChI, and molecular graph-based approaches do support searching and browsing, but their representation of a chemical structure's three-dimensional information is either stored in an ambiguous way or is non-existent. This makes it very difficult for computational chemists to make use of the information found within these approaches as it is the three-dimensional information that is required by many differ-

ent computational methods. On the other hand however, it is very difficult to search for information based on the three-dimensional information alone and as such tools that can support information searching and browsing as well as preserving a chemical structure’s three-dimensional information are needed.

4.1 Multi-Component Data Representation Scheme (MCDRS)

There are many different mechanisms available for capturing the information contained within a chemical structure. In addition to the mechanisms that use chemical properties and metadata (Section 2.2.1), molecular graph theory (Section 2.2.2) and the combination of English language constructs with chemical semantics (Section 2.2.3), there are also matrix based formats and numerical invariants that can be derived from the information found within a chemical structure. Some of the more common examples of these include the atom connectivity matrix [53], and the Wiener number [19].

One of the key components of this thesis is the creation and evaluation of a novel method for capturing the information contained within a chemical structure. When creating this novel data representation scheme, certain goals and objectives were decided upon. A summary of these design goals can be seen in Table 4.1. The purpose of the design goals is to influence how the information used to classify chemical structures is created, stored and accessed. Two of the key design goals are that the representation of a chemical structure is not ambiguous, and that no information

(including the three-dimensional shape of the structure) is lost.

A motivating factor for the development of this Multi-Component Data Representation Scheme (MC'DRS) has been the special needs of researchers that work in the area of computational chemistry. Computational chemists have very complex needs, including being able to distinguish between ground-state geometries and transition-state geometries, as well as being able to identify conformers and structural isomers.

Table 4.1: Multi-Component Data Representation Scheme (MC'DRS) design goals.

Design Goals	Notes
Three-dimensional information is preserved.	Ensure that no information is lost.
The information is non-ambiguous.	The information being used in the data representation scheme is clear, concise and does not have multiple interpretations
The process to generate the data representation scheme is not computationally intensive.	Information for the data representation scheme will need to be processed in real time to support searching and browsing.
The information format is portable and easily accessible.	Not platform or operating system restrictive.
The information format is compatible with and easily integrated into computational chemistry and job scheduling systems.	Easy to use and easy for users to adopt into their workflow.
Supports parallel processing and parallel architecture.	Ensure scalability and flexibility as data sizes and processing requirements grow.

The design goals listed in Table 4.1 have been satisfied through the use of a data representation scheme that combines topological, semantic, and computational information as well as standard three-dimensional Cartesian coordinates, and both the

InChI and canonical SMILES descriptors. Figure 4.1 outlines the structure and the components of this data representation scheme, which is referred to as the Multi-Component Data Representation Scheme (MCDRS). Each of the components will be discussed in the following subsections.

XYZ Cartesian Coordinates	Chemical Atom Topological Indices (CATI)	Chemical Bond Topological Indices (CBTI)
Nuclear Repulsion (complete structure)	Atomic Nuclear Repulsion (contribution of each atom)	Origin-Invariant Nuclear Second Moment
Single-Point STO-3G Energy	InChI	SMILES
<i>Optional Metadata</i> (Level of Theory, Calculation Type, Software, Density Matrix, Frequencies)		

Figure 4.1: Multi-Component Data Representation Scheme - Overview.

4.1.1 Chemical Atom Topological Index (CATI)

It is trivial to label the nodes of a graph generated by a chemical structure with atomic numbers. However, this does not provide much insight into the chemical structure as this same information could be obtained by looking at the chemical formula and the connectivity matrix. Furthermore, if one is to consider each element in the periodic table as a word, then we are given at most 100 of these words to describe all the different chemical structures which is insufficient as it is possible to have very different structures with the same formula. Chemical Atom Topological Index (CATI), developed by R.Poirier (unpublished), has been designed to provide an enhanced level of granularity when representing the atoms that are found within a given structure. This enhanced level of granularity can be attributed to two simple mathematical rules, which are used to define a new vocabulary that describes chemical structures.

In designing this standardized vocabulary, two key factors were considered; namely the canonical ordering of information, and computer processing requirements. By creating rules that do not require any canonical ordering, some of the difficulties that are faced by other methods used to identify and name chemical structures (e.g. SMILES and InChI) have been avoided. Likewise, by building on four simple components: atomic number, valency, maximum valency, and connectivity, the requirement for fast computer processing times has not been compromised. Throughout the development and testing of in-house software associated with this vocabulary it was observed that the average time required to obtain and calculate all of the required information from a Cartesian coordinate file was less than 1 second per structure when using a

2.6GHz computer processor.

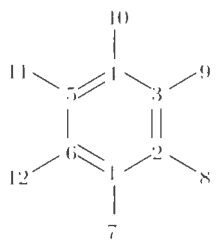
The CATIs are created for each atom by combining the atomic number with the results from two different calculations. The first value, ζ , which uses the connectivity of the atom, is calculated as

$$\zeta = \sum_{i=1}^n (Z - 2)(i) \quad (4.1)$$

where Z is the atomic number. In the calculation of ζ , the atoms are ordered from 1 to n such that their atomic numbers are sorted from lowest to highest. The second value, ξ , is calculated as $(v - v_{max})$ where v is the current valence (# of atoms to which the current atom is bonded) and v_{max} is the maximum valence. The Z , ξ and ζ values are combined to form a CATI, $Z(\zeta, \xi)$.

Figures 4.2 and 4.3 are examples of how the CATI can be used to distinguish two different chemical structures with the same formula. The structure in Figure 4.2 has six atoms with a CATI value of 1(4,0) and six atoms with a CATI value of 6(19, -1). Whereas the structure in Figure 4.3 has six atoms with a CATI value of 1(4,0), five atoms with a CATI value of 6(19, -1), and one atom with a CATI value of 6(9, -1).

Although the CATIs are quite descriptive in nature, they are not unique. It is possible, as it is with words, to have a CATI that has multiple meanings. One example is a carbon atom that is connected to two hydrogen atoms and two oxygen atoms, which has a CATI of 6(39,0). However, a carbon atom that is connected to two other carbon atoms, a nitrogen atom, and a hydrogen atom also has a CATI of 6(39,0).



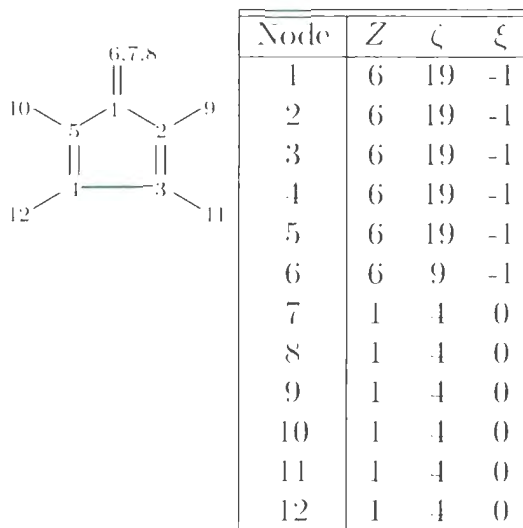
Node	(Z)	ζ	ξ
1	6	19	-1
2	6	19	-1
3	6	19	-1
4	6	19	-1
5	6	19	-1
6	6	19	-1
7	1	4	0
8	1	4	0
9	1	4	0
10	1	4	0
11	1	4	0
12	1	4	0

Figure 4.2: CATI for C_6H_6 example 1.

4.1.2 Chemical Bond Topological Indices (CBTI)

CATI descriptors only describe part of the topological information contained within a chemical structure. If only the CATI were known, and not the connectivity, then it would be difficult to identify important characteristics within the structure. Furthermore, as described in Section 4.1.1, there are cases where CATI can have multiple meanings, and consequently the information associated with CATI is not enough to distinguish a structure from another or to determine if a certain functional group is present or not.

An example of the inadequacy of CATI alone to distinguish some functional groups can be seen in the following example. An oxygen atom that is double bonded to a



The diagram shows a benzene ring with nodes 1 through 12. Nodes 1, 2, 3, 4 form the bottom ring, and nodes 5, 6, 7, 8 form the top ring. Nodes 9, 10, 11, 12 are attached to nodes 2, 1, 3, 4 respectively. Double bonds are shown between nodes 1-2, 2-3, 3-4, 4-1, 5-6, 6-7, 7-8, 8-5. The table to the right provides CATI values for each node.

Node	Z	ζ	ξ
1	6	19	-1
2	6	19	-1
3	6	19	-1
4	6	19	-1
5	6	19	-1
6	6	9	-1
7	1	4	0
8	1	4	0
9	1	4	0
10	1	4	0
11	1	4	0
12	1	4	0

Figure 4.3: CATI for C_6H_6 example 2.

carbon atom would have a CATI of 8(4,-1), and an oxygen that is bonded to both a carbon and a hydrogen would have a CATI of 8(7,0). Unless information is known about how these atoms are connected, and to what, it is impossible to determine whether or not a certain functional group is present. In the case of an alcohol functional group, there is an -OH group bonded to a carbon. However, if that same carbon is also double bonded to an oxygen, then it is considered to be a carboxyl group instead of an alcohol.

By combining the different CATI in such a way that their connectivity and bond information is also captured, additional information about the structure can be stored. Chemical Bond Topological Indices (CBTI), developed by R.Poirier (unpublished), can be defined using CATI and various textual representations of the bonds. The general format of each component of the CBTI representation is as follows:

$$(\text{node}_1 : \text{node}_2)[\text{CATI}_1 < \text{textual bond identifier} > \text{CATI}_2]$$

where (node₁ : node₂) refers to nodes (or atoms) within the structure that are bonded by the bond type indicated in the textual bond identifier, and have CATI given by CATI₁ and CATI₂ respectively. Table 4.2 outlines the different bond types and textual representations used within the CBTI descriptor.

Table 4.2: Bond representation examples, as used in the Chemical Bond Topological Indices (CBTI).

Bond Type	Representation
single bond	—
double bond	=
triple bond	#
aromatic bond	~
weak bond (e.g. transition state - forming, breaking)	*

An example of a CBTI would be (2 : 1) 8(4, -1) = 6(34, -1) which describes a CATI of type 8(4, -1) that is connected by a double bond to a CATI of type 6(34, -1). Figure 4.4 provides an example as to how the combination of CATI and CBTI can be used to identify different components within a particular chemical structure. The first structure in Figure 4.4 contains a carboxyl group which can be identified from the following CBTI:

$$(3 : 1)[8(7, 0) - 6(34, -1)], (3 : 4)[8(7, 0) - 1(6, 0)], (2 : 1)[8(4, -1) = 6(34, -1)]$$

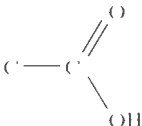
From this information it can be determined that there is a double bond between an oxygen and a carbon (nodes 2 and 1), and a single bond between an oxygen (node 2) that is bonded to a hydrogen and the same carbon (node 1). Using similar infor-

mation, it can also be determined that the second structure in Figure 4.4 only has an -OH group connected to the central carbon and that it is not a carboxyl. The alcohol functional group (-OH) can be identified using CATI descriptors on their own, but the identification of the carboxyl group requires both CATI and CBTI descriptors. This example illustrates how the combination of CATI and CBTI descriptors can be used to distinguish between different functional groups.

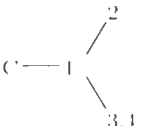
In some cases it is difficult to determine the type of bond that exists between two atoms. When using the Cartesian coordinates of a chemical structure to determine the different types of bonds that are present, the distance between the atoms provides the necessary insight as how best to define a particular bond. For the creation of the CBTI descriptors, the bond order [54] is calculated by evaluating the wave function (using the Hartree-Fock method [55] and the STO-3G basis set [56]). However, if a set of Cartesian coordinates has been incorrectly constructed, or if a three-dimensional structure has been “flattened” to two dimensions, then there is a high probability that some of the bonds will be incorrect. In this case, the Cartesian coordinates would have to be reconstructed or an approximation made by creating a representation of the structure using some type of a chemical editor.

(a) Carboxyl CBTI Example

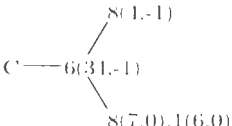
Group



Atomic Number



CATI



Node 1	Node 2	CATI 1	Bond	CATI 2
3	1	8(7,0)	-	6(34,-1)
3	1	8(7,0)	-	1(6,0)
2	1	8(1,-1)	=	6(34,-1)

(b) Alcohol CBTI Example

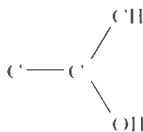
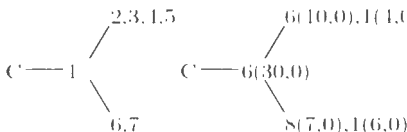
Group	Atomic Number	CATI		
				
				
Node 1	Node 2	CATI 1	Bond	CATI 2
6	1	8(7,0)	-	6(30,-1)
6	7	8(7,0)	-	1(6,0)
1	2	6(30,0)	-	6(10,0)
2	3	6(10,0)	-	1(4,0)
2	4	6(10,0)	-	1(4,0)
2	5	6(10,0)	-	1(4,0)

Figure 4.4: Chemical bond topological indices (CBTI) example differentiating between a structure that has a carboxyl functional group (a) and a structure that has an alcohol functional group (b).

4.1.3 Nuclear Repulsion

The nuclear repulsion and origin-invariant nuclear second-moment (Hollett et al [57] - Section 4.1.4) are properties that are easily calculated using the nuclear coordinates of a chemical structure. Since these properties do not rely on either the electronic wave function or density, they can be calculated quickly. It was observed that on average both the nuclear repulsion and origin-invariant nuclear second-moment can be calculated for a single structure in approximately 0.1 seconds when using a 2.6GHz computer processor for structures with an average molecular complexity [58] of 402.

The following equation is used to calculate the total nuclear repulsion energy (V_{NN}) for a given structure.

$$V_{NN} = \sum_{A=1}^{M-1} \sum_{B=(A+1)}^M \frac{Z_A Z_B}{\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}} \quad (4.2)$$

where M is the number of nuclei within the structure, x_A , y_A , z_A are the Cartesian coordinates of nucleus A, and Z_A is the atomic charge of nucleus A, x_B , y_B , z_B are the Cartesian coordinates of nucleus B, and Z_B is the atomic charge of nucleus B. Within a chemical structure there are three types of interactions, namely the interactions between the nuclei, the interactions between the electrons, and the interactions between the nuclei and the electrons. The nuclear repulsion energy is the energy associated with the interaction between the nuclei.

The total nuclear repulsion can be partitioned into contributions from each atom.

where

$$V_{AtomicNN(A)} = \sum_{B \neq A}^M \frac{Z_A Z_B}{\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}} \quad (4.3)$$

The nuclear repulsion contribution for each atom, calculated using Equation 4.3, provides additional information about a particular atom (e.g. describing an atom's proximity to other atoms within the same structure).

4.1.4 Origin-Invariant Nuclear second-moment

The origin-invariant nuclear second-moment (\hat{S}_N) of a chemical structure is calculated by

$$\hat{S}_N = \begin{bmatrix} \langle \tilde{x}^2 \rangle & \langle \tilde{x}\tilde{y} \rangle & \langle \tilde{x}\tilde{z} \rangle \\ & \langle \tilde{y}^2 \rangle & \langle \tilde{y}\tilde{z} \rangle \\ & & \langle \tilde{z}^2 \rangle \end{bmatrix} \quad (4.4)$$

where

$$\langle \tilde{x}\tilde{y} \rangle = \sum_{A=1}^M Z_A x_A y_A - \frac{1}{N} \left(\sum_{A=1}^M Z_A x_A \right) \left(\sum_{A=1}^M Z_A y_A \right) \quad (4.5)$$

and N is the value of the total nuclear charge ($\sum_A Z_A$), (x_A, y_A, z_A) is the location of A and Z is its atomic charge.

After the diagonalization of \hat{S}_N , the three resulting values (X^2 , Y^2 , and Z^2) cor-

respond to the shape of the molecule. The calculation of the origin-invariant nuclear second-moment is based on a methodology similar to the one outlined by Hollett et al [57], and provides a standardized measure of the shape of a molecule that can be calculated very quickly.

4.1.5 Single Point HF/STO-3G Energy

The calculation of the total energy of a chemical structure at a specific geometry (or single point) can prove to be a useful piece of information, information that can be calculated relatively quickly when compared to the time and computational resources required to calculate an optimized geometry for a given structure. This calculation is more complex than the nuclear repulsion energy calculation as the total energy of a chemical structure is being calculated (i.e. including all types of interactions), and not just a component of the structure's total energy. The total energy value used by the Chem-DRSM system is approximated using the Restricted Hartree-Fock (RHF) method with the STO-3G basis set.

Although the Single Point HF/STO-3G energy not necessarily associated with a fully optimized geometry or stationary point, the information obtained from this calculation can be particularly useful when used as a criteria for comparing chemical structures.

4.1.6 InChI and SMILES descriptors

As described in Section 2.2.3, the InChI and SMILES descriptors are very useful and have become industry standards. By including both the InChI and SMILES chemical

descriptors within the data representation scheme, compatibility with existing chemical information systems can be ensured. Furthermore, users that are used to working with InChI and SMILES representations can easily transition their work to use some of the other components of the data representation scheme that are better suited to their needs.

4.1.7 Metadata / Additional Properties

The design of the data representation scheme has made provisions for the inclusion of meta or descriptor-type tags and keywords. Also included in the data representation scheme is the ability to store experimentally and computationally derived properties. Although these fields are currently treated as basic textual fields, the design and flexibility of the data representation scheme allows for these fields to be fully indexed and used by the various similarity metrics without impacting any of the existing functionality.

Examples of some of the properties and metadata descriptors that can be included are: the software used to complete the geometry optimization calculation, the level of theory (energy approximation method and basis set), any citations related to the chemical structure, drug company index numbers or references, drug company vendor sources, cost, calculated frequencies, and the density matrix. These examples are not exclusive as other properties may be included; they only serve to illustrate the diversity and functionality available within this data-representation scheme.

4.2 Implementation and design of the Chem-DRSM system

Building on the Multi-Component Data Representation Scheme, the next step is to incorporate this scheme into the design of a similarity metric (or set of similarity measures) that will allow for fast, accurate, and reliable searching of chemical structures using information contained within the data representation scheme. Table 4.3 outlines the design goals for the search and retrieval component.

The Chem-DRSM system provides a solution that builds on the Multi-Component Data Representation Scheme and satisfies the design goals outlined in Table 4.3. This system automatically processes the three-dimensional information contained within a chemical structure, and proceeds to generate suitable indices and data representations. Once this information has been stored, the indices and the information contained within the data representation scheme can be used to assess chemical similarity and search for important sub-structures or components (i.e. functional groups). The remainder of this chapter discusses the components and design of the Chem-DRSM system.

The Chem-DRSM system has been implemented using a series of interconnected modular components. A conceptual design drawing of the Chem-DRSM system can be seen in Figure 4.5 where the complete system is shown along with the intermediary chemical structure representations that are produced. The subsections pertaining to the Chem-DRSM system have been grouped according to their placement within the

Table 4.3: Similarity engine design goals.

Design Goals	Notes
Exact chemical structures can be matched	This feature would be required by many users.
Ability to distinguish chemical structures based on their conformation.	Conformation searching is difficult because of how the three-dimensional information is represented in commonly used data representation schemes.
Ability to search for substructures and functional groups.	These features are important as it allows for flexible searching and browsing, providing not only a measure of overall similarity but also localized similarity.
Ability to support both Boolean and ranked queries.	Both query types would allow for greater flexibility when meeting the needs of users.
Ability to support bulk operations.	Allow for the integration and automatic processing of large data sets.
Ability to support individual queries.	Allow for interactive work to be completed.
Ability to perform a single query in a reasonable amount of time (on par with current search engine technology).	Performance targets are important. If the system is too slow, then it would not be used.

architectural design of the system as a whole. The first subsection reviews all the components relating to the pre-processing of the chemical structure data. The second subsection discusses the information extraction process, the third subsection discusses the index creation process and the fourth subsection describes the chemically based similarity measures that are part of the Chem-DRSM system.

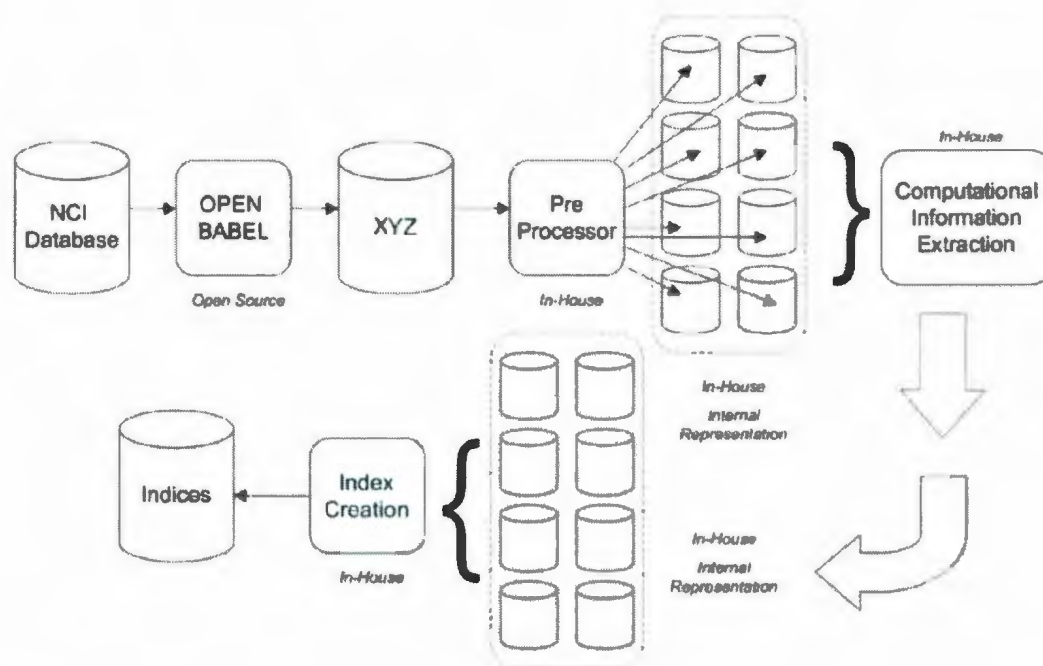


Figure 4.5: Modular architecture of the Chem-DRSM system.

4.2.1 Structure Pre-Processing

Chemical structure information is pre-processed prior to information extraction for two reasons. First, to ensure that the chemical structure has valid three-dimensional information. Second, to ensure that the chemical information is in a form that can be processed by the information extraction component. The information extraction component is designed to read in either Cartesian coordinates or Z-matrix formats. The Cartesian coordinate format was specifically chosen as it is supported as a valid output format by OpenBabel [14, 15], and many different file formats (e.g. sdf, mol, Gaussian-03 [59] input and output files) can be easily converted to the Cartesian coordinate representation using OpenBabel. Figure 4.6 provides an example of the Cartesian coordinate representation and the sdf file representation for $C_3H_4O_2$. Note that the three-dimensional information is not altered, only the representation and the formatting of the data has been changed.

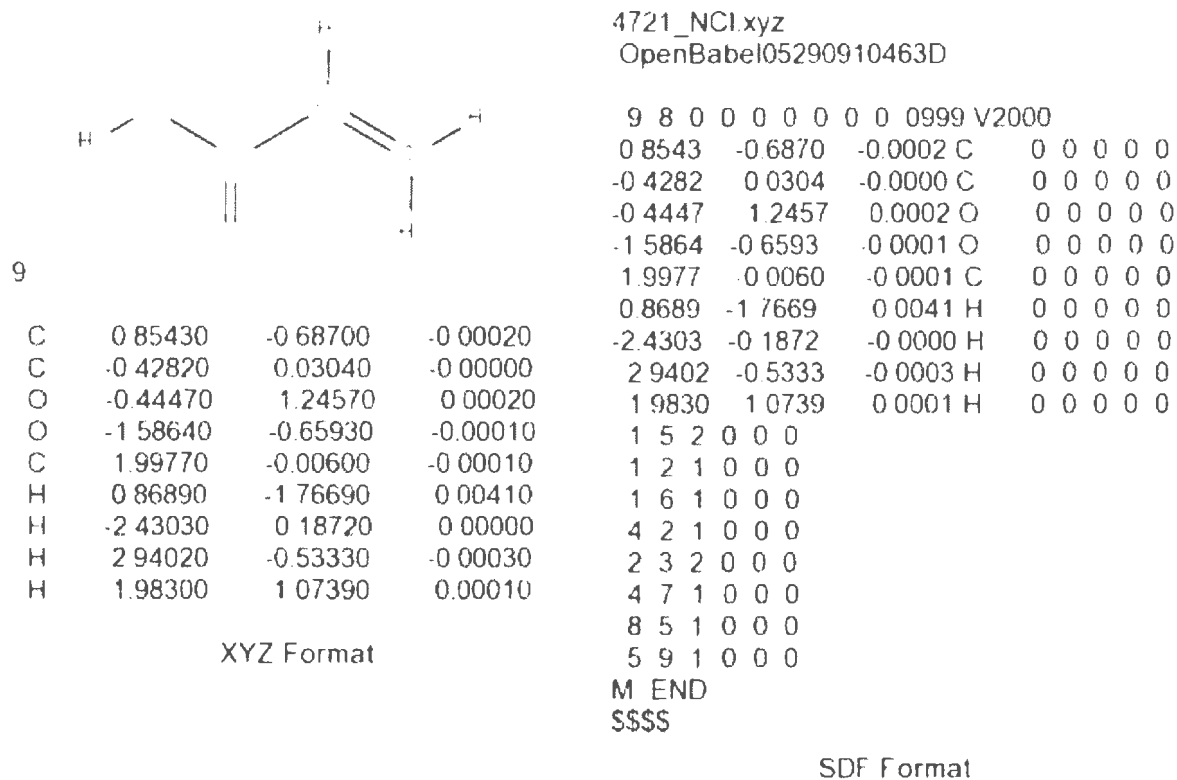


Figure 4.6: Example Cartesian coordinate and sdf representations for $C_3H_4O_2$, as well as a molecular graph of the structure for comparison. Note, the drawing of the structure is only included to aid the reader and is not part of either the SDF or XYZ data files.

4.2.2 Information Extraction

A key part of the information extraction process has been to use the computational chemistry software package MUNgauss [60] which has the ability to process the information contained within the Cartesian coordinate representation of a chemical structure. Only a small subset of MUNgauss’s functionality is used for determining the information that is required for similarity purposes. These components, when combined with the translation abilities of OpenBabel, are referred to as the computational information extraction component in Figure 4.5. The computational information extraction component then uses information contained within the chemical structure to calculate the CATI, CBTI, and computationally derived descriptors (nuclear repulsion, origin-invariant nuclear second-moment, and single point energy), as well as metadata and industry standard descriptors, all of which are components found in the Multi-Component Data Representation scheme, as shown in Section 4.1. Table 4.4 shows all of the different components extracted from a given Cartesian coordinate representation of a chemical structure, and the modules used within the Chem-DRSM system.

Table 4.4: Information derived from the Cartesian coordinate representation of a chemical structure by the Chem-DSRM system.

Property	Component Used for Extraction
CATI (Topological Descriptor)	MUNgauss
CBTI (Topological Descriptor)	MUNgauss
Nuclear Repulsion (Computationally Derived Descriptor)	MUNgauss
Origin-Invariant Nuclear Second-Moment (Computationally Derived Descriptor)	MUNgauss
STO-3G Single Point Energy (Computationally Derived Descriptor)	MUNgauss
Canonical SMILES (Industry Standard Descriptor)	OpenBabel
InChI (Industry Standard Descriptor)	Open Babel
Chemical Formula (Metadata Descriptor)	MUNgauss
NCI Index Number (Metadata Descriptor)	Chem-DSRM Translator / Builder (only if included in source data)

4.2.3 Index Creation

Upon the completion of the information extraction process, each chemical structure will have a number of different descriptors associated with it. The use of indices to further organize these descriptors allows for the easy integration of many different types of search tools and similarity measures. For every descriptor two different indices are required, namely an index that links the structure to the descriptor, and an index that links the descriptor to one or more structures. Figure 4.7 provides an example of the two indices that are produced for the canonical SMILES descriptor. The Chem-DRSM system creates the following indices to support searching and browsing activities, and to provide a consolidated means to use the information found within key descriptors:

- CATI \longleftrightarrow Structure Indices
- Nuclear Repulsion \longleftrightarrow Structure Indices
- Single Point Energy \longleftrightarrow Structure Indices
- Origin-Invariant Nuclear Second-Moment \longleftrightarrow Structure Indices
- Formula \longleftrightarrow Structure Indices
- NCI Number \longleftrightarrow Structure Indices
- Canonical SMILES \longleftrightarrow Structure Indices
- InChI \longleftrightarrow Structure Indices

Canonical SMILES	# of Structures	Structure List
<chem>CSc1cccc1C(=O)O</chem>	5	109910,113808,145573,146505,81687
<chem>CSCC[C@H](N)C(=O)O</chem>	5	203719,21022,39716,9092,90849
<chem>COc1cc(/C=C\C(=O)C)ccc1O</chem>	5	24042,38997,3985,39482,5252
<chem>CCCCCCCC(=O)OC=C.C=CN1CCCC1-O</chem>	5	87737,87738,87739,87740,87741
<chem>CC(C)O[P@](=O)(C)O</chem>	5	164544,164581,164586,164682,164731
<chem>CC(=O)CC[C@@H]1CC(=O)OC1(C)C</chem>	5	106812,106824,32216,39673,98243

Structure ID	Canonical SMILES
93688	<chem>Cl/C=C(\c1ccc(Cl)cc1)/c1cccc1Cl</chem>
8563	<chem>CCOc1ccc(cc1)C(=O)O</chem>
58017	<chem>CCCCCCCCCCCC[C@@H](O)CO</chem>
44919	<chem>Cl.OCCN(CCO)Cc1nc2cc(Cl)c(Cl)cc2[nH]1</chem>
127219	<chem>o.oC1(CCCCC1)c5c1cccc1</chem>
87220	<chem>Cl.Oc1ccc2CCNCCc2c1</chem>
63406	<chem>CCCCCCCCNCCO</chem>
55548	<chem>CCC\C(=C(/CC)\C#N)\C</chem>
31922	<chem>CC1=CC(=O)C(=CC1=O)C</chem>
18581	<chem>CCCCCOc(=O)C(CC)CC</chem>
67525	<chem>O=C(C(=O)Nc1cccc1)c1cccc1</chem>
89241	<chem>O=C1NC(=O)C2(CCNCC2)N1C</chem>
77451	<chem>CCC(=O)c1c(O)cc(O)cc1O</chem>

Figure 4.7: Example of the indices created for the canonical SMILES descriptor. (Note how the two relationships are preserved SMILES \rightarrow Structure, Structure \rightarrow SMILES)

4.2.4 Chemically Based Similarity Measures

As illustrated in Chapter 2, there are many different ways of capturing the information found within a chemical structure. Similarly, Chapter 3 demonstrates that there are many different ways of assessing the similarity of two chemical structures. Figure 4.8 shows the modular design of the similarity measures found within the Chem-DRSM system.

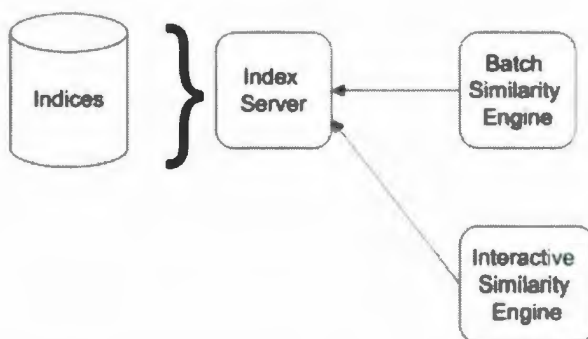


Figure 4.8: Architecture of the similarity measures found within the Chem-DRSM system.

There are two important features found within the design of the similarity measures that are of note. First, it independently uses the indices created by the build and information extraction process, as seen in Figure 4.5. This means that the index building and information extraction processes can be executed without influencing the behaviour of the similarity measures. Second, the similarity measures employ a design that has been influenced by a client-server framework. The client-server

architecture found within the Chem-DRSM system allows different similarity metrics to use the same index data and it allows for system designers and maintainers to take advantage of different scalability and load balancing techniques, an example of which can be seen in Figure 4.9.

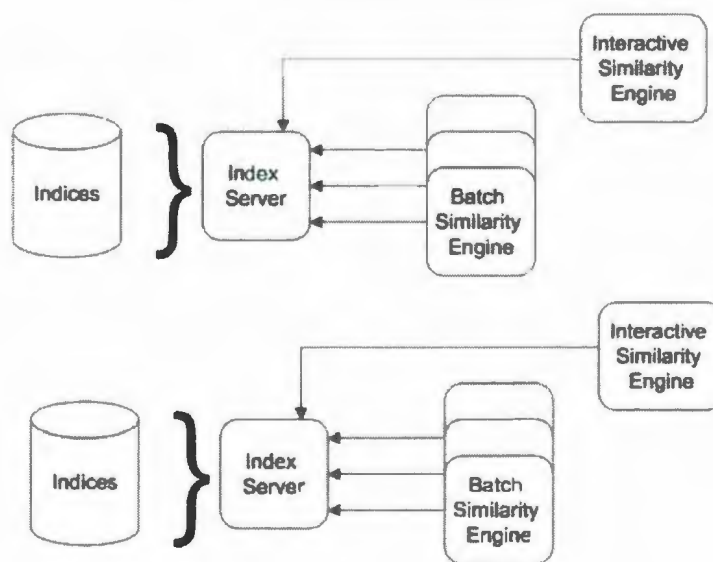


Figure 4.9: Example of similarity measure scalability within the Chem-DRSM system.

The Chem-DRSM system currently supports a number of different similarity metrics that use the information that is contained within the various indices, as itemized in Section 4.2.3. The similarity measures used by the Chem-DRSM system fall into two different categories, namely vector space models, and refinement via computationally derived descriptors.

4.2.4.1 Vector Space Models

Section 3.3 outlines some of the more common approaches for producing similarity scores. The use of the Tanimoto [48] coefficient with chemical fingerprints is a very widely used measure of chemical similarity [61]. As an alternative to the “standard” Tanimoto coefficient, a modified version of the Tanimoto measure has been implemented. The modified Tanimoto measure uses CATI descriptors instead of chemical fingerprints, and has been implemented using a modified version of Equation 3.1, where N_a represents the number of CATI descriptors found in structure A, N_b represents the number of CATI descriptors found in structure B and $N_{a \cap b}$ represents the number of CATI descriptors common to structures A and B.

It can be argued that the Tanimoto coefficient (and other similar similarity coefficients) are limited in how the similarity coefficients are calculated. One characteristic of the Tanimoto similarity coefficient is that it does not take into consideration the significance, or weighting, of each of the structural fragments or properties that make up the structure. In the example shown in Figure 3.2, Chapter 3 one can see that it is possible, depending on the properties being considered, for two chemical structures with different formulas (in this case $C_1H_{10}OS$ and $C_6H_{14}O_2S$) to have a similarity score of 1.0 (100% similarity).

Not considering the statistical distribution of the properties being used in the calculation of the similarity score gives cause for concern as frequently occurring properties are treated the same as rare properties. This is equivalent to having a linguistic search

tool that places the same significance on the term *and* and the term *photosynthesis*.

One way to correct how properties are weighted is to use the cosine measure (see Section 3.3). Although the cosine measure is more computationally expensive than the Tanimoto measure it has gained a wider acceptance as a standard measure within information retrieval circles as it has the ability to capture the context of the terms in a given query. Witten et al [1] describe a document vector as a ray emanating from the origin, piercing space in some desired direction. Extending this description, the task of searching for similar documents can be described as the process of selecting those document vectors that lie closest to the ray in an angular sense. The angle between two document rays, or chemical structure rays, is called θ . The similarity of the two representative vectors can be examined by looking at θ . The cosine of θ equals 1 when $\theta = 0^\circ$, which means that there is no difference in the representative vectors. Additionally, when the cosine of θ equals 0 the vectors are at right angles to each other, which means that the representative vectors are unrelated. By computing the cosine of the angles between the two vectors a similarity score between 0 and 1 is produced.

The cosine measure traditionally uses words or keyphrases found within document texts to produce representative vectors. The Chem-DRSM system draws on this and modifies the cosine measure to use CATI. In this context the CATI can be thought of as chemical "words". Two adaptations of the cosine measure to support the use of CATI have been implemented within the Chem-DRSM system, namely a standard cosine measure and a contextually based cosine measure. These two cosine variants

are commonly used within information retrieval systems, and as the CATI descriptors are a new type of chemical structure descriptor, it was important that both variants be considered.

The two adapted versions of the cosine measure are a standard cosine measure, Equation (4.6), and a contextual cosine measure, Equation (4.7) and are determined as follows:

$$\text{standard cos}(Q, D) = \frac{\sum_{t \in Q \cap D} (f_{t,Q} \cdot \log \frac{n}{f_t})(f_{t,D} \cdot \log \frac{n}{f_t})}{\sqrt{\sum_{t \in Q} (f_{t,Q} \cdot \log \frac{n}{f_t})^2} \sqrt{\sum_{t \in D} (f_{t,D} \cdot \log \frac{n}{f_t})^2}} \quad (4.6)$$

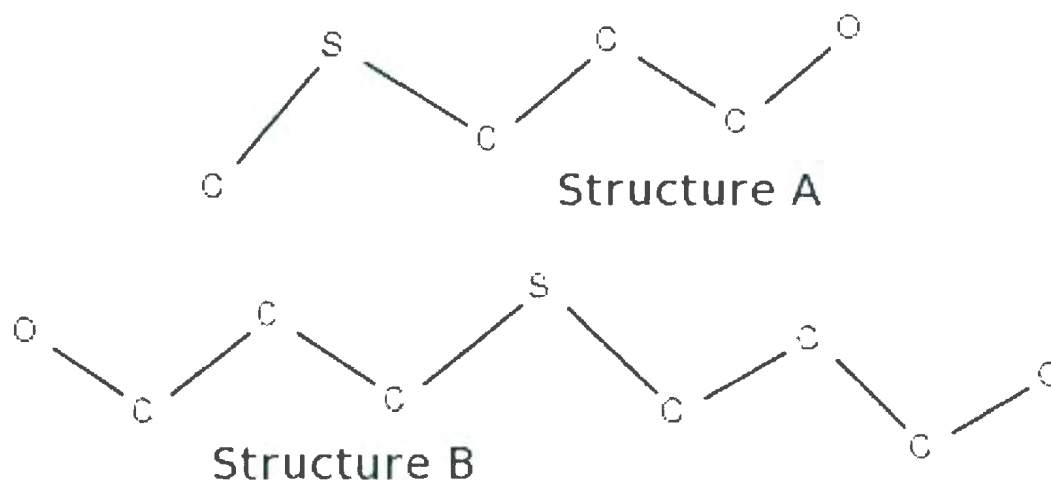
$$\text{contextual cos}(Q, D) = \frac{\sum_{t \in Q \cap D} (f_{t,Q} \cdot \log \frac{n}{f_t})(f_{t,D} \cdot \log \frac{n}{f_t})}{\sqrt{\sum_{t \in Q} (f_{t,Q} \cdot \log \frac{n}{f_t})^2} \sqrt{\sum_{t \in Q} (f_{t,D} \cdot \log \frac{n}{f_t})^2}} \quad (4.7)$$

where Q is the query structure, D is the structure being compared to the query structure, n is the number of structures in the database, f_t is the number of structures that contain CATI t , $f_{t,Q}$ is the number of times CATI t occurs within the query structure, and $f_{t,D}$ is the number of times CATI t occurs within the structure being compared to the query structure. It is worth highlighting that the only difference in the two cosine measures is how the denominator is calculated. The current implementations of these two cosine measures within the Chem-DRSM system uses weighting information that has been derived from over 178,000 different chemical structures from the NCI chemical structure database.

The key difference between the standard cosine measure and the contextual cosine

measure is how the two chemical structure vectors are compared. Each chemical structure can be described as a list of CATI descriptors, as shown in Figure 4.10. When representing a chemical structure, both the types and quantities of the CATI that are present within the structure are considered. The different CATI descriptors can be thought of as vector dimensions, and the quantities of a given CATI within a structure can be used to calculate the magnitude of the vector in that direction.

Figure 4.10 shows an example of both the CATI descriptors and the frequencies for $C_1H_{10}OS$ and $C_6H_{14}O_2S$. When comparing the two structures, the vector space used to determine similarity can be described in different ways. The standard cosine uses the CATI descriptors found in both the query (Q - what is being searched for) and the CATI descriptors found in the document (D , or in this case, chemical structure) being compared. Alternatively, the contextual cosine places a higher significance on the CATI descriptors found within the query, as opposed to the structure being compared. It is important to note that the vector space defined by the contextual method is dependent on which structure is the query. However the vector space defined by the standard cosine is the same for the two structures involved, regardless of which structure is the query and which structure is the comparison.



	16(12.0)	8(7.0)	6(65.0)	6(50.0)	6(33.0)	6(25.0)	1(6.0)	1(4.0)
Structure A	1	1	1	1	1	1	1	9
Structure B	1	2	2	0	2	2	2	12

Figure 4.10: CATI listing, with quantities, for $C_4H_{10}OS$ and $C_6H_{14}O_2S$. (Note: hydrogen atoms are not shown in structural representations.)

In summary, when using the contextual method to define the vector space it is possible for the relationship defined by the comparison of Q to D , to be different than the relationship defined by the comparison of D to Q . However, when using the standard cosine method to define the vector space, the relationship defined by the comparison of Q to D is the same as the relationship that is defined by the comparison of D to Q .

In addition to determining the appropriate vector space model for the cosine measures, there is also the task of determining an appropriate statistical weighting scheme, unlike the Tanimoto coefficient which is determined using information relating to if a certain feature or descriptor is present and not how statistically significant it is. A

common weighting approach used in information retrieval science combines the term frequency (how many times a given term occurs within a document) with the document frequency (how many documents within the entire library (or collection) contain that particular term). This type of weighting approach has recently been adapted by Lipkus et al.[40] to assist with a further analysis of the diversity and composition of the structures within the Chemical Abstracts Service (CAS) registry database.

In the case of the work being done in this thesis, the weighting scheme has been derived from the structures within the NCI database. It is important to note that additional training data may be needed to make this weighting scheme more general in nature. However, being able to create a weighting scheme from given structures would allow the maintainers of different chemical information resources to create their own weighting schemes based on the area of focus of that particular research group or resources.

Although both of these cosine measures can be used to assess the similarity of different chemical structures, there are still instances where these equations cannot distinguish two different structures. It is in these cases where computationally derived chemical descriptors can be used to further assess and refine chemical structure similarity.

4.2.4.2 Similarity refinement using computationally derived chemical descriptors

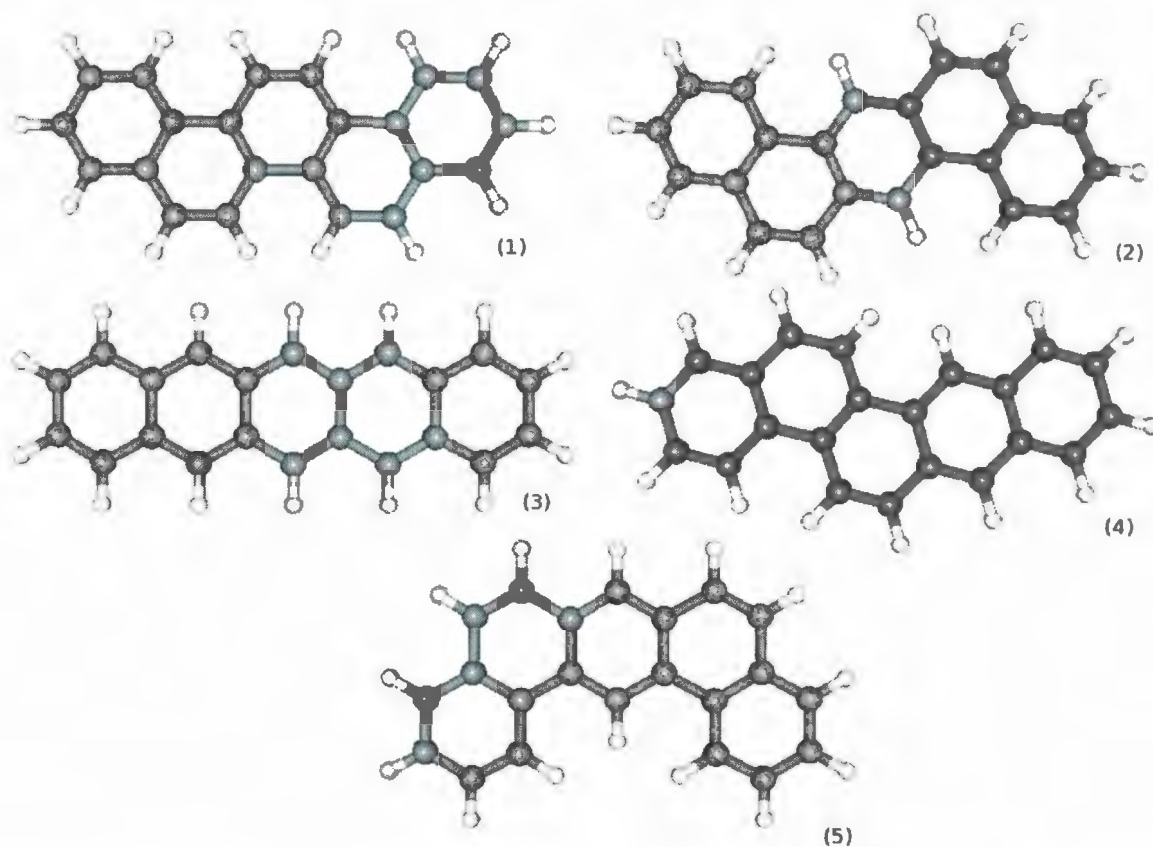
To aid in the assessment and refinement of chemical structure similarity, components of the Multi-Component Data Representation Scheme (see Section 4.1), were eval-

uated to determine their suitability. The values for nuclear repulsion energy (see Section 4.1.3), origin-invariant nuclear second-moment (see Section 4.1.4), and the single point energy of a chemical structure calculated using the STO-3G basis set (see Section 4.1.5) were considered. Computational experiments were conducted to determine the acceptable variances of these three values, which were then used as threshold values for determining and refining chemical structure similarity.

Chapter 5, Section 5.1, outlines the experimental procedure and results that were used to establish thresholds for similarity refinement using these computationally derived chemical descriptors. By comparing the nuclear repulsion energies, the origin-invariant nuclear second-moment, and STO-3G single point energy values of over 20,000 chemical structures it was observed that there is a 2.16% variation in nuclear repulsion energy, a 5.23% variation in the origin-invariant nuclear second-moment, and a 0.00178% variation in the single point STO-3G energies when comparing optimized geometries that have been completed using the same structure, just different basis sets.

Figure 4.11 shows five structures with the same chemical formula, $C_{22}H_{14}$, obtained from the NCI database. These structures cannot be differentiated by their chemical formula, or by the use of CATI or CBTI descriptors. However, by comparing the values in Figure 4.11 with the computed thresholds, the structures can be differentiated. In terms of the origin-invariant nuclear second-moment values, there are some structures that have values that fall within the 5.23% threshold, but none of the structures have all three values (X^2 , Y^2 and Z^2) falling within the 5.23% threshold of another

structure. Similarly, there are only two structures (1 and 5) out of the five structures that have STO-3G energy values that fall within the 0.00178% threshold of the other structures. None of the $C_{22}H_{11}$ structures shown in Figure 4.11 matched for all values (nuclear repulsion, origin-invariant nuclear second-moment, and STO-3G single point energy) with another structure. This example shows promise for the use of these computationally derived properties to allow for further refinement of candidate lists of similar and / or matching chemical structures.



	Nuclear Repulsion Energy (Hartrees)	Origin-Invariant Nuclear Second-Moment (Bohr ²)			Single Point Energy RHF / STO-3G (Hartrees)
		X ²	Y ²	Z ²	
1	1548.1600	0.0015	1071.8667	6398.0813	-831.0702
2	1536.8993	0.0002	1145.4508	6427.2661	-842.2723
3	1504.3097	0.0012	815.6531	8008.2693	-831.0082
4	1536.5591	0.0004	1039.0668	6688.8879	-831.0552
5	1545.6834	0.0004	1538.6471	5550.2588	-831.0734

Figure 4.11: Example illustrating different nuclear repulsion, origin-invariant nuclear second-moment and single point energy values (RHF/STO-3G) for different C₂₂H₁₄ structures.

Chapter 5

Investigative Approach and Results

This chapter discusses observations and results from four different investigations that related to the Chem-DRSM system. The first section of this chapter describes the experimental process used to determine the similarity thresholds of three different computationally derived properties, namely nuclear repulsion energy, origin-invariant nuclear second-moment, and single point energy as calculated using the STO-3G basis set. The second section of this chapter presents a statistical evaluation of the different metrics within the Chem-DRSM system (standard cosine, contextual cosine, and Tanimoto with CATI descriptors) and compares them to a Tanimoto measure that makes use of Chemical Fingerprints. Building on the statistical evaluation, the third section presents results from a human study that compares the same Chem-DRSM metrics with the Tanimoto Chemical Fingerprint measure, however in this case the comparison is based on the assessments made by 24 study participants with expert knowledge in Chemistry. The final section of this chapter presents an additional investigation that was completed in order to determine the suitability of the CATI and CBTI descriptors to rapidly identify the presence of different functional groups within a chemical structure.

5.1 Determination of Similarity Threshold for Computationally Derived Descriptors

As introduced in Chapter 4, Section 4.2.4.2, there is a need to be able to further screen the search results that are produced by the various metrics found within the Chem-DRSM system that make use of CATI descriptors. Three computationally derived descriptors, namely nuclear repulsion energy, origin-invariant nuclear second-moment, and the single point energy of a chemical structure as calculated using the STO-3G basis set, were chosen as candidates to perform further similarity screening.

An experiment was conducted using a sample collection of 877 different optimized structures from the computational and theoretical chemistry research group at Memorial University. These structures (stored as MUNgauss archive files) ranged in size and composition, and contained anywhere from zero to twelve carbon atoms. The idea behind the experiment was to use different basis sets while optimizing the chemical structure geometries to produce different optimized geometries for the same initial chemical structure. It was expected that the variances within the different basis sets, would be observed as similarity thresholds for the different computationally derived descriptors.

The first stage of the experiment was completed by optimizing the geometry of each of the 877 structures using each of five different basis sets: 3-21G, STO-3G, 6-31G, 6-31G(d), and 6-31+G(d), and the restricted Hartree-Fock (RHF) method. This

produced five different optimized geometries for each of the 877 structures (4385 structures in total). The 4385 resulting structures were then each optimized again using the RHF method and the five basis sets listed above. This resulted in 25 new optimized geometries for each of the 877 initial structures (21,925 in total). Figure 5.1 outlines all of the possible combinations produced for a single structure, and note that there are five different optimized final geometries for each of the basis sets that were tested.

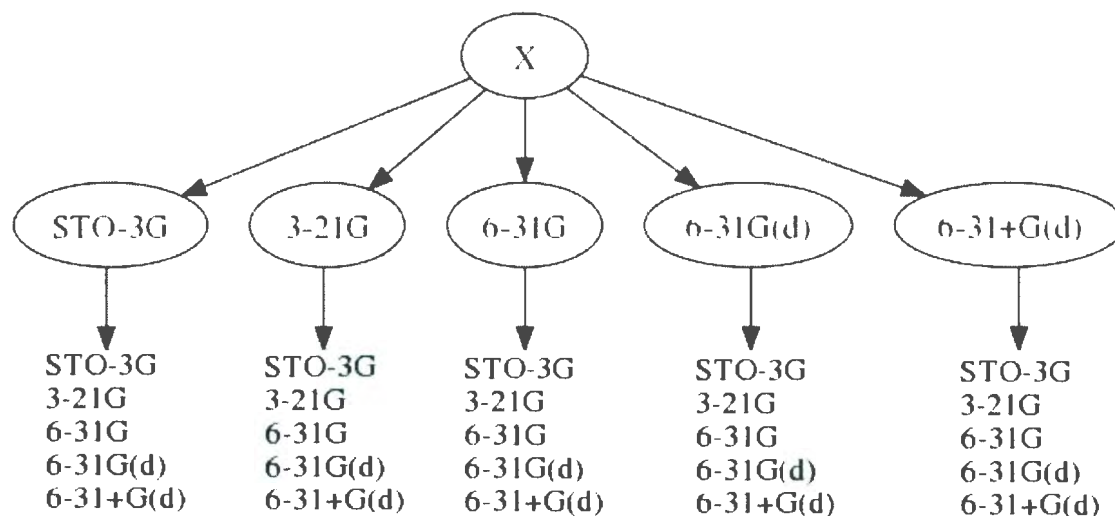


Figure 5.1: Calculation methodology showing the different optimized geometries that are used to determine the nuclear repulsion threshold, the origin-invariant nuclear second-moment threshold, and STO-3G single point energy threshold for any given structure (X).

Through basic statistical analysis it became possible to determine the average variance in nuclear repulsion energy, origin-invariant nuclear second-moment, and single point energy values for each of the optimized geometries that were derived from each

of the 877 initial structures. The 25 different optimized geometries for each structure provides a maximum, minimum and variance value for each of the computational descriptors. Taking the average variance value, as a percentage, for each of the 877 groups shows that there is a 2.16% variation in nuclear repulsion energy, a 5.23% variation in the origin-invariant nuclear second-moment, and a 0.00178% variation in the single point STO-3G energies when comparing optimized geometries that have been completed using the same structure, just different basis sets (Table 5.1). As demonstrated in Chapter 4, Section 4.2.4.2, these threshold values can be applied to sets of chemical structures to further distinguish similar structures. A further area where these thresholds can be of use is when trying to distinguish between structures that are conformers or structural isomers.

Table 5.1: Calculated thresholds for nuclear repulsion, origin-invariant nuclear second-moment, and STO-3G single point energy values.

Nuclear Repulsion	Origin-Invariant Nuclear second-moment	STO-3G Single Point Energy
2.16%	5.23% (X^2, Y^2 , and Z^2)	0.00178%

•

It is important to note however, that the calculated thresholds are to be considered an upper limit as there were observed cases where the resulting geometry optimizations are conformers of the original structure. The presence of these conformers within the analysis population inflates the threshold values. Manual screening of the $\sim 25,000$ structures within the analysis population would serve to reduce the threshold values.

5.2 Statistical Evaluation

The statistical evaluation component is designed to simulate the use of the Chem-DRSM system by conducting searching and browsing activities through the creation of ranked lists of chemical structures from a test collection. This has important potential applications, as there are millions of chemical structures stored on computers throughout the world, and if these structures could be accessed using a system that was more comprehensive in nature, then great benefits could be seen in how researchers search for and access chemical structures.

The statistical experiment was carried out using the following procedure:

- Chemical structures were converted to Cartesian coordinate format.
- Chemical Atom Topological Indices (CATI) were determined for each chemical structure in the test collection.
- Indices were created based upon the resulting CATI descriptors.
- 19 query structures were chosen.
- The similarity of each chemical structure within the test collection (as it compared to each of the 19 query structures) was determined through the use of the indices and three different similarity measures, namely the standard cosine measure, the contextual cosine measure, and the Tanimoto CATI measure (as implemented within the Chem-DRSM system).

- The similarity of each chemical structure within the test collection (as it compared to each of the 19 query structures) was also determined through the use of a Tanimoto similarity metric that uses chemical fingerprints (as implemented within the OpenBabel system).
- Listings of the different structures with a similarity score of 1.0 (as produced by the three Chem-DRSM metrics and the Tanimoto chemical fingerprint metric) were produced for each of the 19 query structures (four different metrics, 19 different structures, 76 different lists of chemical structures)
- The different listings of the structures that had scores of 1.0 were evaluated statistically (using precision and recall).
- Histograms were created to further evaluate the nature of the similarity scores produced by the different measures.

The following subsections describe each of these items in more detail.

5.2.1 Choosing the test collection

The test collection of chemical structures was made up from structures found in the National Cancer Institute (NCI) online database (Release 3 Files - September 2003) [3]. Out of the 260,071 structures found within the NCI database 178,175 structures were found to have suitable three-dimensional information. The remaining 81,896 structures were not converted because they were stored as two-dimensional projections of the three-dimensional information found within the original structure (which resulted in a loss of original information).

This collection was selected for a number of reasons, primarily because the data associated with the 260,071 structures was available free of charge, but also because the structures in the collection are quite diverse and representative of structures that a wide range of chemists would use.

5.2.2 Information extraction and index creation

Using in-house software that calculates the CATI descriptors and determines the quantities of the different CATI descriptors within each structure, summary files for each of the 178,175 chemical structures in the test collection were produced. Based upon the values contained within the summary files for each chemical structure, the required indices for the Chem-DRSM system (as outlined in Section 4.2.3) were created.

5.2.3 Chemical structure similarity computations

In order to determine the similarity between different chemical structures, three different in-house similarity measures were employed. To assess these three similarity measures with respect to a baseline measurement, the standard Tanimoto metric with chemical fingerprints (as implemented by OpenBabel) was also used. A summary of the measures used for the evaluation can be seen in Table 5.2. The three in-house similarity measures are explained in detail in Section 4.2.4 and more detail about the Tanimoto metric that uses chemical fingerprints can be found in Section 3.3.

Table 5.2: Methods that were used to produce similarity scores.

Method 1	Method 2	Method 3	Method 4
cosine (Contextual + CATI)	cosine (Standard + CATI)	Tanimoto (CATI)	Tanimoto (chemical fingerprints)
Chem-DRSM Equation (4.7)	Chem-DRSM Equation (4.6)	Chem-DRSM Equation (3.1)	OpenBabel Equation (3.1)

5.2.4 Precision and Recall Evaluation

The first component of the evaluation of the different chemical similarity metrics involves a statistical evaluation. The statistical evaluation was conducted using two standard measures, namely precision and recall.

The precision (P) of a similarity measure for some cutoff point (r) is the fraction of the top r ranked items that are relevant to the query.

$$P = \frac{\text{number retrieved that are relevant}}{\text{total number retrieved}} \quad (5.1)$$

For example, if one hundred chemical structures are retrieved in response to a particular query ($r = 100$), and fifty of them are relevant, then the precision of the similarity measure would be 50%. The precision metric measures the accuracy of the search.

Complementing this is the recall measure. The recall measure (R) for a particular r value (some cutoff point) is the proportion of the total number of relevant items retrieved within the top r .

$$R = \frac{\text{number relevant that are retrieved}}{\text{total number relevant}} \quad (5.2)$$

Continuing the example used with the precision measure (where $r = 100$), if there are seventy five relevant items in the entire collection then the recall of the similarity measure would be 66% since only fifty out of the seventy five were selected. Recall measures the extent to which the retrieval is exhaustive and quantifies the coverage of the items that are returned.

The biggest difficulty with this type of evaluation is identifying a standard set of documents, queries, and relevance judgements (decisions as to which documents in the collection are answers to each query). In this experiment, a subset of the NCI database was used as the standardized document set, and both the queries and answers to the queries were obtained from within this collection. To obtain relevance judgements, a combination of canonical SMILES and InChI representations of the chemical structures were used. A chemical structure was considered to be a "correct answer" to a query if it had both the same canonical SMILES and the same InChI representation as the structure used for the query.

5.2.5 Data Analysis

Using ranked lists as obtained from the four different similarity measures, recall and precision values were determined. Since recall is a nondecreasing function of rank (its position in the list), precision can be regarded as a function of recall rather than rank. Moreover, precision is usually high at low recall levels and low at high recall levels, if one were to plot a precision-recall curve, the curve generally decreases. If a perfect ranking algorithm could be developed, all relevant items would be ranked

ahead of all irrelevant items. In this case, precision would be 100 percent at all recall levels, and the recall-precision curve would be a horizontal line at 100 percent.

For the 19 structures selected to be part of the statistical evaluation, the r (or cutoff value, as relating to both precision and recall) used was a score of 1.0 as opposed to a set numerical threshold (e.g. first ten structures returned). Since the order of the returned structures was dependent on the order in which they were placed into the index, rather than some other descriptive property, the information found within a recall-precision curve would be biased as to their index placement. Instead the precision and recall values were calculated by reviewing all of the structures with a score of 1.0.

In order to compute all the similarity scores for the 19 different structures with the four different similarity measures, over 13 million pairwise similarity calculations were required to be completed (4 million with OpenBabel and 12 million with the ChemDRSM system). Table 5.3 lists the formulas, cancer chemotherapy National Service Center (NSC) numbers, number of structures that are identical, and the number of structures with the same formula within the test collection for each of the 19 different structures used in the statistical evaluation.

All four of the similarity measures had 100% recall with each of the 19 test structures, and it was only through the use of the precision results that any difference between the four similarity measures was observed. Table 5.4 shows the precision results for the four different similarity measures for each of the 19 test structures.

Table 5.3: Formulas, NSC numbers, number of structures that are identical, and the number of structures with the same formula in the test collection for the 19 different structures used in the statistical evaluation.

NSC ID	Formula	Number of identical structures with same formula	Number with same formula
131564	$C_6H_{10}N_2O_2$	5	43
134438	$C_9H_{16}O_3$	8	56
152324	$C_2H_6O_2$	14	14
169899	$C_{18}H_{19}NSCl_2$	5	5
170347	$C_8H_8O_4$	5	54
209826	$C_5H_{12}N_2O_3$	7	12
210746	$C_8H_{11}N_4O_2$	7	18
1880	$C_9H_{10}O_3$	5	98
79367	$C_7H_8O_4$	6	30
4765	$C_3H_4O_2$	8	15
8134	$C_{10}H_{11}O_3$	3	54
90799	$C_8H_{17}N$	2	25
134422	C_2H_5N	4	6
131564	$C_6H_{10}N_2O_2$	5	42
26613	$C_{11}H_{12}O_3$	4	68
623441	$C_9H_8O_2$	2	19
153096	C_9H_8OS	2	1
525079	$C_9H_9NO_3$	2	73
15309	$C_8H_8O_2$	2	33
167530	C_6H_6	3	1

When comparing the different similarity measures it was observed that, on average, the Tanimoto similarity measure with chemical fingerprints had a precision of 75% and a standard deviation (σ) of 31%. This is in contrast to the Chem-DRSM measures which had an average precision of 92% with a standard deviation of 17% for the standard cosine measure, an average precision of 70% with a standard deviation of 37% for the contextual cosine measure, and an average precision of 66% with a standard deviation of 38% for the Tanimoto measure with CATI descriptors.

Although the average precision and standard deviation values provide some insight into the behaviour of the different measures, the results do not give a complete picture. For example, the average precision of the standard cosine measure is 92% and the standard deviation is 17%. Although this shows the magnitude of the deviation, it also makes one question how can there be a precision of 109% ($92 + 17$)? To obtain more insight additional statistical summary data was determined for the different data sets (as can be seen in Table 5.4). The range of the data points shows that the values found within the standard cosine data set are closer together than those found in the other data sets. Furthermore, the skew (a measure of the lack of symmetry of a distribution curve, a distribution curve is symmetric if it looks the same to the left and right of the center point) and the kurtosis (a measure of the slope of the data curve, the value of which describes if the data curve is peaked or flat as compared to a normal distribution) values are very different for the standard cosine measure than those found with the other three measures. The skew value shows that the distribution of the data points for the standard cosine are much more condensed towards the right hand side of a distribution curve (negative skew values correspond to a data

distribution where the left tail of the distribution is longer than the right tail of the distribution and positive skew values correspond to a data distribution where the right tail of the distribution is longer than the left tail of the distribution). and the kurtosis value shows that the curve associated with the standard cosine measure has a sharper peak (positive kurtosis indicates a sharper peak in the data distribution, whereas a negative kurtosis value indicates a flatter distribution).

The shape of the distribution curves provides additional information about how to interpret the standard deviation and the average precision. In the case of the standard cosine measure, the skew and the kurtosis demonstrate the the majority of the data points are concentrated around the middle of the curve (92% precision) and that the data points with the greatest influence on the standard deviation are on the left hand side of the curve (less than 92%).

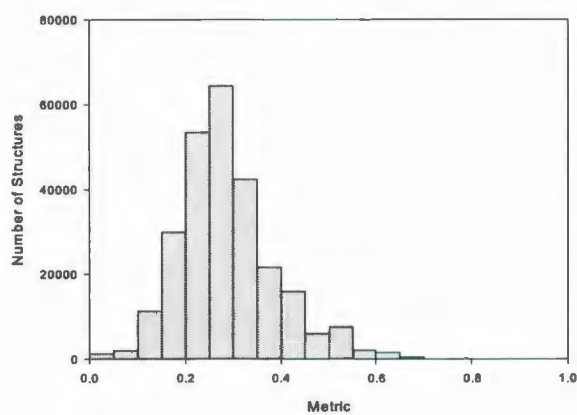
Although these results are only for a small number of structures from within the entire test collection (19 query structures or 99 structures if you include exact matches), the results still show that within the scope of the study, the standard cosine measure has the smallest range of precision values, the highest average precision value and the smallest standard deviation of precision values. These values show the validity of the approach being taken by the Chem-DRSM system with the CATI based measures.

To assist with further analysis, the statistical evaluation was extended beyond exact matches. When considering the quality of results produced by similarity measures it is important to consider the ability of the similarity measure to appropriately score

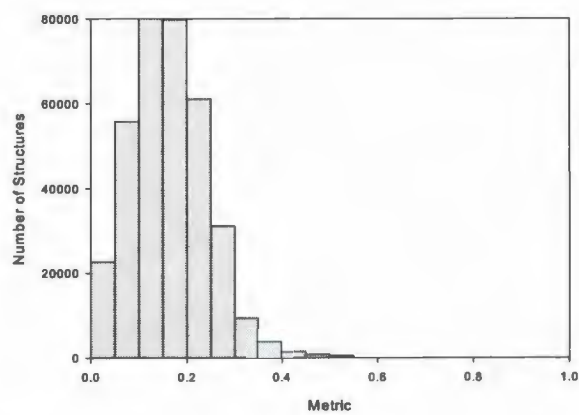
Table 5.4: NSC numbers, precision values and statistical summary data for the 19 structures that were part of the statistical evaluation.

NSC ID	Precision (%) Tanimoto (chemical fingerprints) Chem-DRSM Equation (3.1)	Precision (%) cosine (Standard + CATF) Chem-DRSM Equation (4.6)	Precision (%) Tanimoto (CATF) Chem-DRSM Equation (3.1)	Precision (%) cosine (Contextual + CATF) OpenBabel Equation (4.7)
131564	100	100	100	100
134422	100	57	13	57
134438	100	100	100	100
152324	100	100	25	100
153096	50	100	67	29
167530	23	75	0	75
4765	100	100	50	29
169899	100	100	100	100
170347	42	100	100	83
209826	100	100	100	100
210746	88	88	88	88
15309	12	40	17	1
1880	71	100	83	56
525079	100	100	100	67
623441	100	100	8	3
26613	80	100	100	7
79367	50	100	100	100
8134	30	75	75	18
90799	53	100	67	100
Average	73.61	91.32	63.84	68.05
Median	87.50	100.00	75.00	83.00
Mode	100.00	100.00	100.00	100.00
Range	88	60	99	100
Deviation (σ)	30.2	16.8	36.8	36.1
Skew	-0.72	-2.09	-0.56	-0.78
Kurtosis	-0.99	3.77	-1.32	-1.04

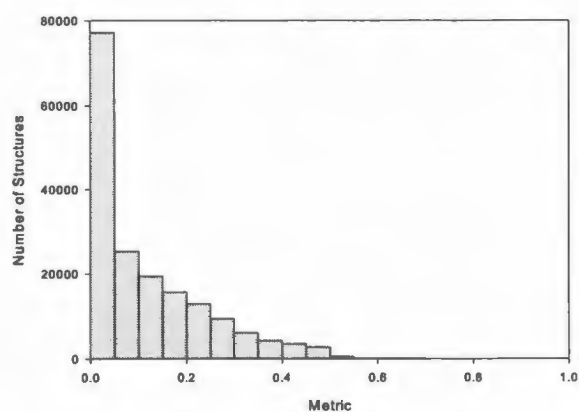
differences, both large and small, between the different items being considered. For example, a measure that simply looks at the presence and absence of features instead of their statistical significance and quantities might have abrupt changes in similarity scores, whereas a metric that is more granular in nature can detect more subtle differences and make appropriate scoring adjustments. Insight into the behaviour of the metric with respect to its granularity can be seen in the histograms of similarity scores produced for various queries. Figures 5.2 through 5.5 show histograms of the results produced by the four different similarity measures with four of the query structures from the statistical evaluation (NSC' 131564, NSC' 134422, NSC' 134438, and NSC' 152324) which provide representative data for all of the histograms produced by the 19 query structures. The complete histogram data for each of the four measures and the 19 query structures with the test collection (over 13 million similarity scores) can be reviewed in Appendix A. Please note that the scales found within the histograms have been kept the same so as to highlight the differences in score distribution between the different metrics. As such, the complete data for a given column may not be shown.



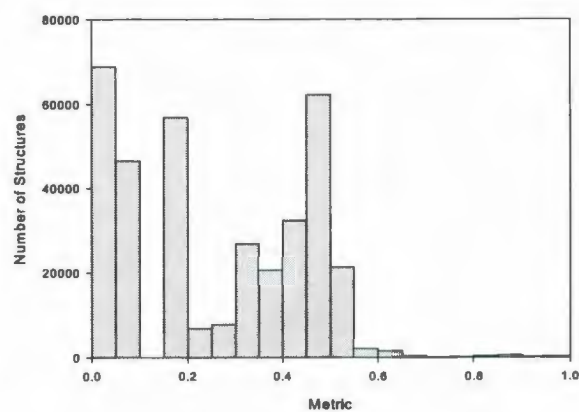
(a) Tanimoto fingerprints



(b) Tanimoto CATI

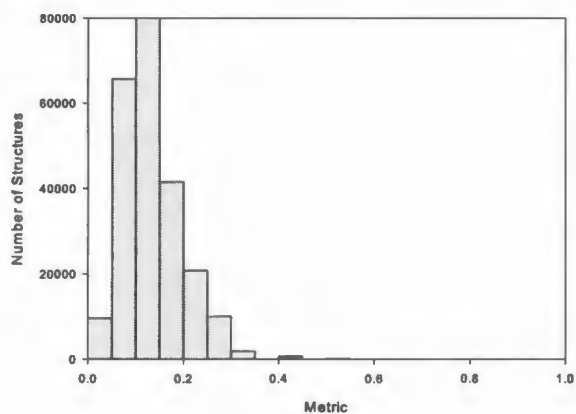


(c) Standard cosine

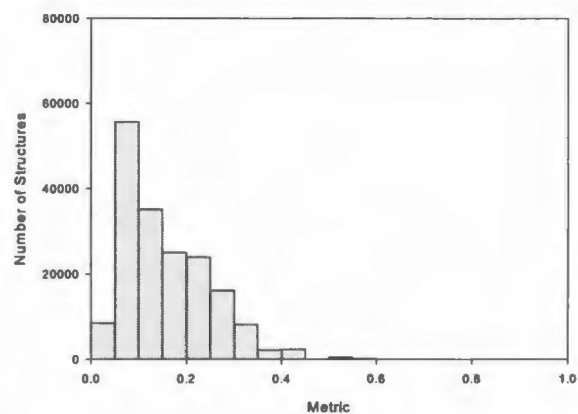


(d) Contextual cosine

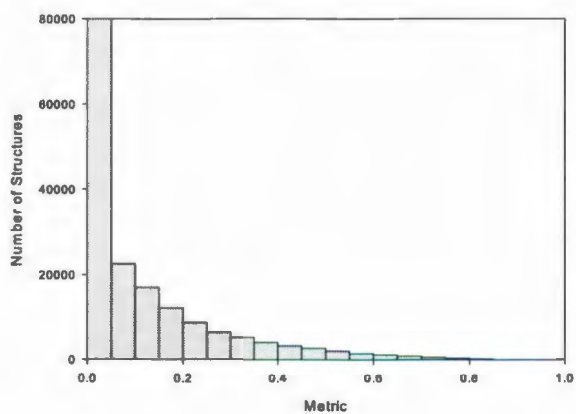
Figure 5.2: Histograms of results (similarity scores) for query structure NSC 131564 with 4 similarity measures (as indicated)



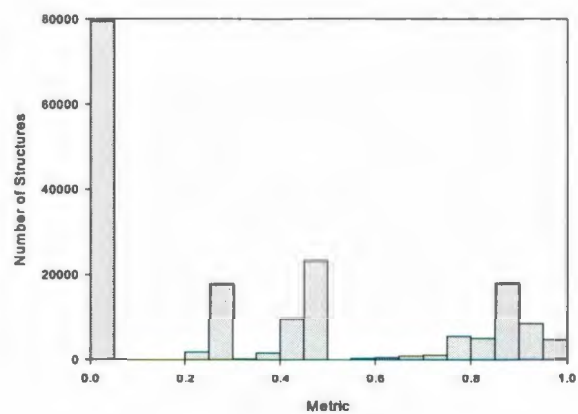
(a) Tanimoto fingerprints



(b) Tanimoto CATI

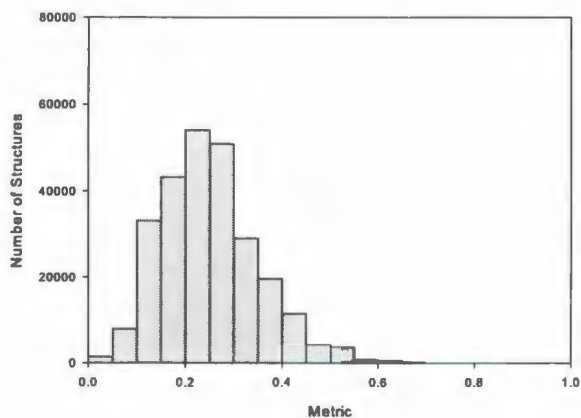


(c) Standard cosine

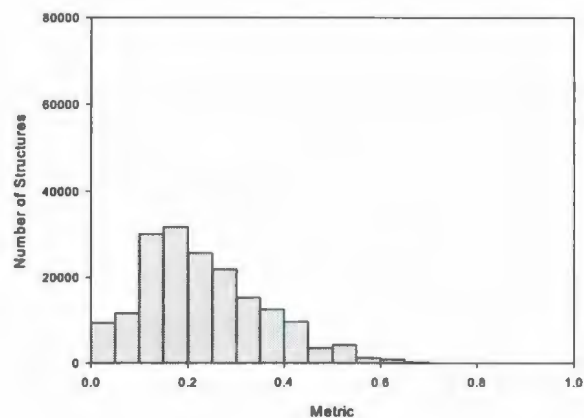


(d) Contextual cosine

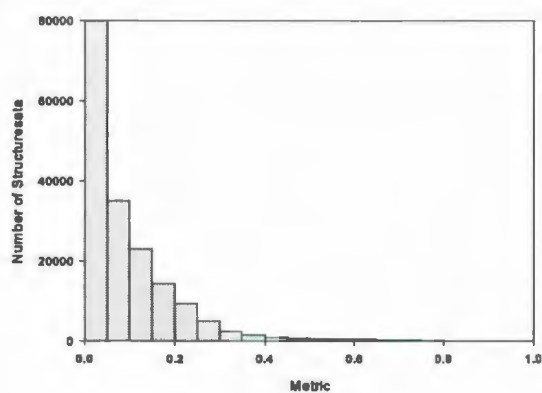
Figure 5.3: Histograms of results (similarity scores) for query structure NSC 134422 with 4 similarity measures (see Table A.2 for data).



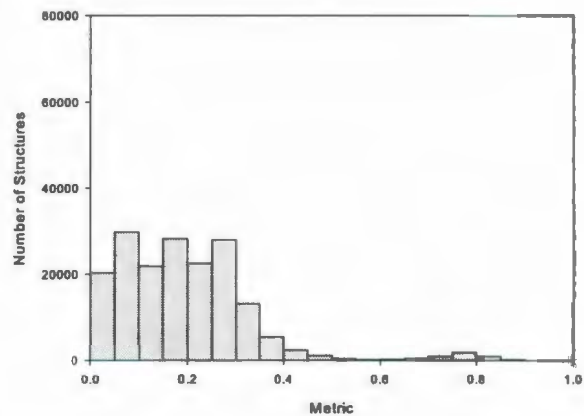
(a) Tanimoto fingerprints



(b) Tanimoto CATI

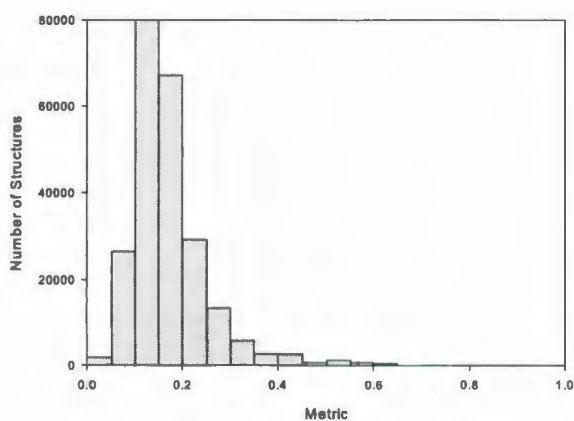


(c) Standard cosine

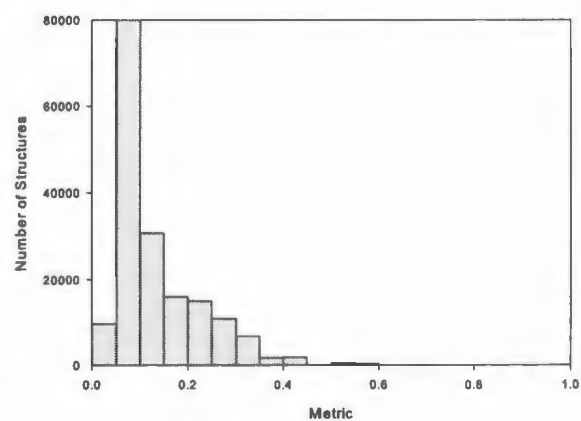


(d) Contextual cosine

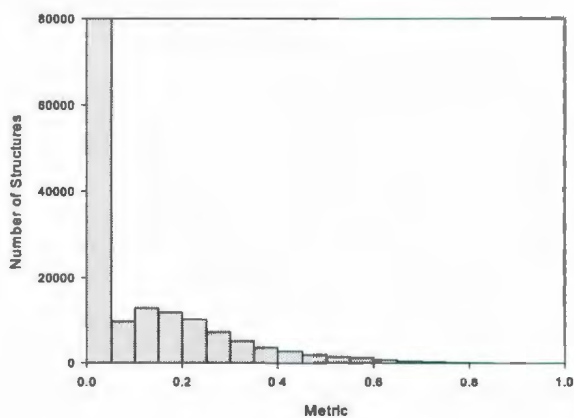
Figure 5.4: Histograms of results (similarity scores) for query structure NSC 134438 with 4 similarity measures (see Table A.3 for data).



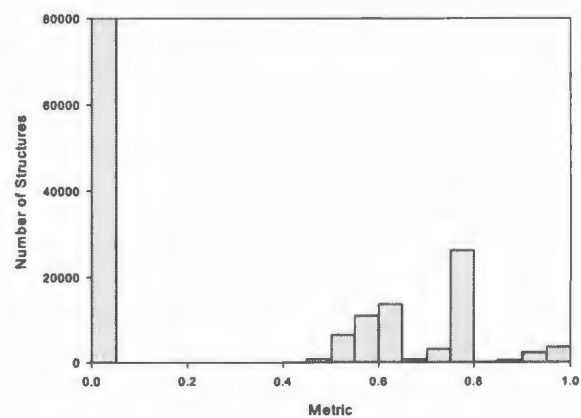
(a) Tanimoto fingerprints



(b) Tanimoto CATI



(c) Standard cosine



(d) Contextual cosine

Figure 5.5: Histograms of results (similarity scores) for query structure NSC 152324 with 4 similarity measures (see Table A.4 for data).

As can be seen from the histograms, the different metrics behave differently when assessing similarity. If the metrics were similar in nature, then the histograms would either be the same or linear translations of each other. Since this is not the case, observations can be made from the histogram data that allows for further differentiation of the metrics. Consider Table 5.5, which shows the histogram data for the one of the 19 query structures (NSC 90799, C_8H_{17}). This structure is of interest when reviewing the distribution of scores produced by the standard cosine measure and the Tanimoto measure with chemical fingerprints because of the distribution of the similarity scores. When the standard cosine measure is used, a score of 100% precision is observed for exact matches (Table 5.4), and the progression through the similarity scores of the entire test collection is done gradually as the chemical structures become more and more dissimilar to the query structure (3 structures with scores between 1.0 and 0.96, 58 structures with similarity scores between 0.91 and 0.95, and 462 structures with similarity scores between 0.86 and 0.90). This is in contrast to the Tanimoto measure with chemical fingerprints, which in addition to having a precision of 53% when searching for exact matches (Table 5.4) has 22 structures with scores between 1.0 and 0.96, 0 structures with scores between 0.81 and 0.95, and 117 structures with scores between 0.76 and 0.80. The large spread of values with 0 scores (0.81 - 0.95) shows the lack of granularity in the Tanimoto measure, behaviour which is also shown in Table 5.6 (NSC 167530 - C_6H_6) when the Tanimoto measure with chemical fingerprints is in use. In this case there are 26 structures with scores between 0.96 and 1.0, 0 structures with scores between 0.56 and 0.95, and 327 structures with scores between 0.51 and 0.55. This is in contrast to the standard cosine measure which, in this case, has 5 structures with scores between 0.96 and 1.0, 5 structures with scores

between 0.91 and 0.95 and 39 structures with scores between 0.86 and 0.90. Upon further investigation of the query results for each of the 19 query structures, it was determined that the Tanimoto measure with chemical fingerprints has, on average, 3.75 histogram data ranges (0.00 to 0.05, 0.06 to 0.10, 0.11 to 0.15, 0.16 to 0.20, 0.21 to 0.25, 0.26 to 0.30, 0.31 to 0.35, 0.36 to 0.40, 0.41 to 0.45, 0.46 to 0.50, 0.51 to 0.55, 0.56 to 0.60, 0.61 to 0.65, 0.66 to 0.70, 0.71 to 0.75, 0.76 to 0.80, 0.81 to 0.85, 0.86 to 0.90, 0.91 to 0.95, and 0.96 to 1.0) per structure with values equal to zero, whereas the standard cosine measure, on average, has 0.1 histogram data ranges per structure with values equal to zero.

The presence of a similar trend within the results from the Tanimoto measure that uses the CATI descriptor (average of 3.55 histogram data ranges per structure with values equal to zero) makes one conclude that the block-like behaviour (similar to a step function) of the histogram results observed within the results produced by the Tanimoto measure that uses chemical fingerprints is more likely to be attributed to the nature of the Tanimoto equation (for example the statistical distribution) rather than the chemical fingerprint descriptor. This observation re-iterates the importance of taking into account the statistical distribution and weighting of the components being used to determine similarity.

Table 5.5: Distribution of similarity scores produced by different similarity measures when structure NSC 90799, $C_8H_{17}N$, is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	5	3	8	22
0.95	5517	58	0	0
0.90	6557	462	0	0
0.85	25842	1263	0	0
0.80	8986	1642	43	0
0.75	5901	1698	0	117
0.70	2199	1838	99	0
0.65	5696	2082	18	11
0.60	2633	2393	232	383
0.55	2406	2708	509	0
0.50	692	3310	928	1198
0.45	7483	3963	895	0
0.40	12488	4543	2818	2936
0.35	2988	5245	4649	2259
0.30	637	6291	10520	9999
0.25	380	7600	19849	23963
0.20	20544	9376	28068	37846
0.15	3210	11858	33159	51309
0.10	14452	15512	29915	84594
0.05	44355	22218	37646	38690
0.00	5204	74112	8819	6744

Table 5.6: Distribution of similarity scores produced by different similarity measures when structure NSC 167530, C_6H_6 , is the Query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (1.7)	cosine (Standard + CATI) Chem-DRSM Equation (1.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	13729	5	5	26
0.95	36683	5	0	0
0.90	11637	39	0	0
0.85	6014	112	0	0
0.80	2400	246	0	0
0.75	1860	423	0	0
0.70	1719	773	0	0
0.65	1283	1023	131	0
0.60	1017	1460	0	0
0.55	696	1938	0	0
0.50	565	2605	385	327
0.45	443	3722	0	0
0.40	430	4893	1309	0
0.35	483	6751	0	0
0.30	185	9508	2699	1165
0.25	71	12797	14572	2773
0.20	13	16427	28557	5035
0.15	0	21056	48247	8381
0.10	39723	23299	52009	65194
0.05	0	17949	21449	96668
0.00	59224	53144	8812	80502

5.3 Human Evaluation

Although the statistical analysis, as described in Section 5.2, attempted to obtain insight into the behaviour of the different similarity measures, it can be thought of as incomplete since opinions of potential users of the Chem-DRSM system were not yet considered. To effectively compare the three proposed similarity metrics with the Tanimoto measure that uses chemical fingerprints, a study was designed to complement the statistical analysis. This study involved the use of human subjects to evaluate and score the pairwise similarity scores and the correctness of the list rankings that were produced by the three proposed similarity metrics and the Tanimoto metric with chemical fingerprints.

The human evaluation of the different similarity measures was carried out using the following procedure:

- Chemical structures were converted to Cartesian coordinate format.
- Chemical Atom Topological Indices (CATI) were determined for the each chemical structure in the test collection.
- Indices were created based upon the resulting CATI descriptors.
- Five query structures were identified for use with the human evaluation.
- The similarity of each chemical structure within the test collection (as it compared to each of the five human study query structures) was determined through the use of the indices and three different similarity measures: namely the stan-

dard cosine measure, the contextual cosine measure, and the Tanimoto CATI measure (as implemented within the Chem-DRSM system).

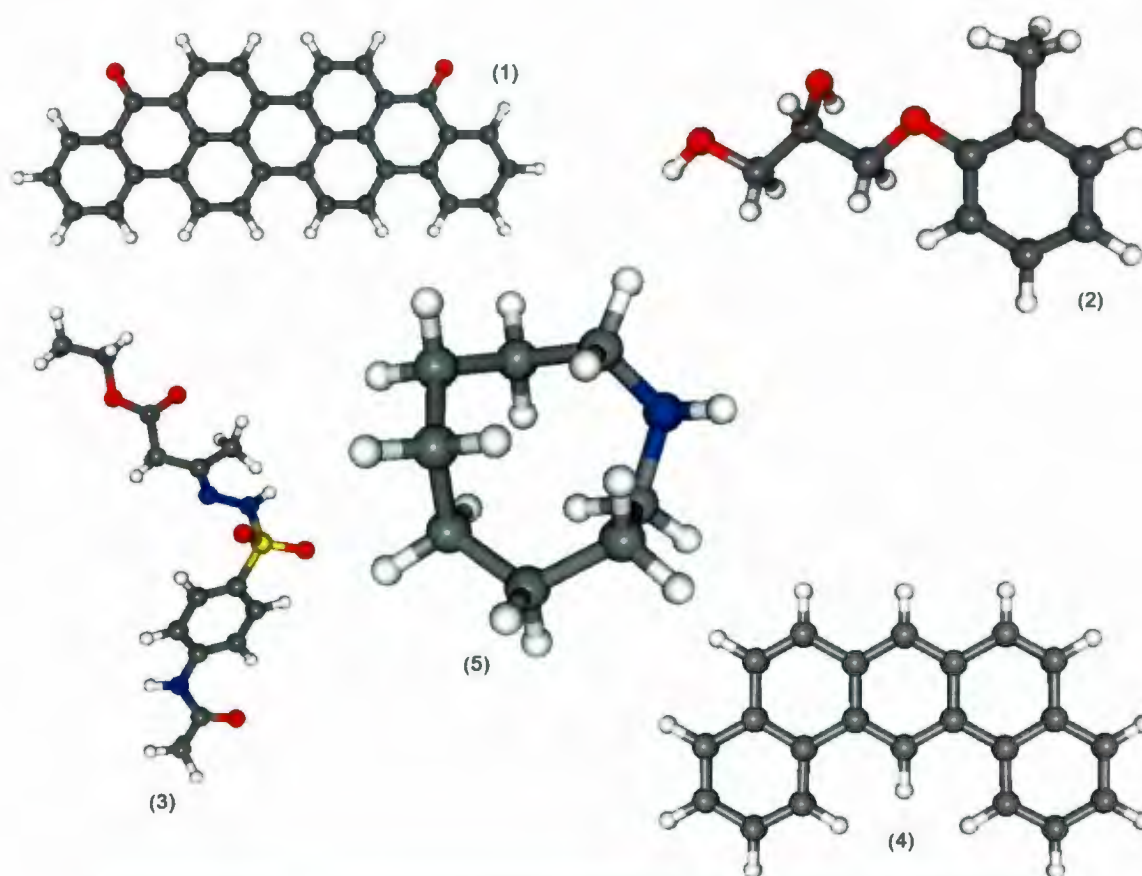
- The similarity of each chemical structure within the test collection (as it compared to each of the five human study query structures) was also determined through the use of a Tanimoto similarity metric that uses chemical fingerprints (as implemented within the OpenBabel system).
- Lists of the top ten most similar structures (as determined by the three Chem-DRSM metrics and the Tanimoto chemical fingerprint metric) were produced for each of the five human study query structures (four different metrics, five different query structures, 20 different top ten lists of chemical structures). In cases where there were more than ten structures with a score of 1.0, only the first ten structures returned were used.
- The different top ten listings were evaluated in terms of pairwise similarity and correctness of list ordering by the human study participants.

The following subsections describe each of these items in more detail.

5.3.1 Building on the Statistical Evaluation

The human evaluation builds on the methodology and framework that was used in the statistical analysis Section 5.2. The same test collection, index creation scheme and similarity measures as the statistical evaluation were also employed in the human evaluation. The only difference in the experimental foundation is that the human evaluation only looks at the query results produced by five different query structures

from the test collection whereas the statistical evaluation used 19 different query structures from the test collection. Figure 5.6 shows the structures, formulas and NSC identification numbers of the five different chemical structures that were arbitrarily chosen for use with the human evaluation.



1	2	3	4	5
NSC 2212	NSC 8134	NSC 18084	NSC 90321	NSC 90799
$C_{34}H_{16}O_2$	$C_{10}H_{14}O_3$	$C_{14}H_{19}N_3O_5S$	$C_{22}H_{14}$	$C_8H_{17}N$

Figure 5.6: Query structures that were used to produce similarity scores for the human evaluation (the choice was arbitrary).

5.3.2 Hypothesis

The main focus of the study was to determine the appropriateness in scoring and ordering of the results from the various chemical structure queries. It was the intent of this study to find the answer to the following question, “Is there a difference in the quality of results that are returned by the different similarity metrics?”. To answer this question, the following null hypothesis was used: “There is no difference in the similarity scores produced by the four different similarity metrics that are being considered in terms of precision, recall, distribution of similarity scores, and user assesment.”

5.3.3 Subjects

Twenty-four university students, staff and faculty members were recruited for this experiment. Tables 5.7, 5.8 and 5.9 show the demographic background of the subjects that participated in the study. This study not only provided an evaluation of the different similarity metrics, but it also provided some additional insight into what types of criteria the study participants used to determine chemical structure similarity. Prior to the commencement of the study, ethical approval was sought and granted by the Interdisciplinary Committee on Ethics in Human Research (ICEHR) at Memorial University of Newfoundland.

Table 5.7: Gender breakdown of study participants.

Male	Female
17 (71%)	7 (29%)

Table 5.8: Educational background of study participants.

BSc	MSc	MSc	PhD
(Chemistry)	(Chemistry)	(Computational Science)	(Chemistry)
1 (4.2%)	3 (12.5%)	1 (4.2%)	19 (79.2%)

Table 5.9: Area of expertise of study participants.

Number of study participants	Area of expertise
2 (8.3%)	Theoretical
6 (25%)	Theoretical / Computational
4 (16.7%)	Physical
2 (8.3%)	Physical / Computational
1 (4.2%)	Physics / Condensed Matter
3 (12.5%)	Organic / Experimental
2 (8.3%)	Organic
1 (4.2%)	Crystallography / Inorganic
1 (4.2%)	Inorganic
2 (8.3%)	Analytical

5.3.4 Method

Each subject was required to complete two tasks for each of the different similarity measures. The first task consisted of scoring the similarity of structures within a list of chemical structures to a structure that was designated as the search query. The second task involved providing an overall score for the correctness of the ordering of the list that was produced. For each task, there were five lists generated using the test collection from the NCI database (NSC 18084 - $C_{11}H_{19}N_3O_5S$, NSC 2212 - $C_{31}H_{16}O_2$, NSC 8134 - $C_{10}H_{11}O_3$, NSC 90321 - $C_{22}H_{14}$, and NSC 90799 - $C_8H_{17}N$) and the generation of each of the lists required $\sim 178,000$ similarity calculations. The mechanics of the four different tasks are all the same, it is simply the content of the

generated lists that are different.

To account for the possibility that the results could be influenced by the order in which the tasks were completed, counterbalanced measures were implemented (see Table 5.10). The 24 subjects were each given a different ordering of tasks so that all of the possible task orderings were considered. The only variation within the study was the ordering of the different similarity measures. Throughout the experiment subjects were presented the same lists for each of the similarity measures in the same order. Table 5.11 illustrates the ordering of the structures used to produce the lists of chemical structures that were reviewed by each study participant.

It is important to note that each of the lists reviewed by the study participants contained only ten structures. A total of ten structures was chosen as the cutoff for user evaluation, as a study by Beitzel et. al. [62] using web-based queries from more than 50 million users showed that users only view the results presented on the first page (first ten documents) of a web-query 81% of the time. Since the purpose of this evaluation is to simulate and evaluate searching and browsing activities only the first ten structures were presented to study participants, even if there were more than ten with the same similarity score.

5.3.5 Data Collection

Throughout the course of the experiment data relating to the following areas was collected: similarity to the original query structure and correctness of list ordering. In order to measure both of these values a scale from one to seven, otherwise known

Table 5.10: Summary of similarity measure ordering for subject tasks.

	cosine (Contextual + CATI)	cosine (Standard + CATI)	Tanimoto (CATI)	Tanimoto (chemical fingerprints)
Subject A	1	2	3	4
Subject B	1	2	4	3
Subject C	1	3	2	4
Subject D	1	3	1	2
Subject E	1	4	2	3
Subject F	1	4	3	2
Subject G	2	1	3	4
Subject H	2	1	4	3
Subject I	2	3	1	4
Subject J	2	3	4	1
Subject K	2	1	1	3
Subject L	2	4	3	1
Subject M	3	1	2	1
Subject N	3	1	1	2
Subject O	3	2	1	4
Subject P	3	2	1	1
Subject Q	3	1	1	2
Subject R	3	4	2	1
Subject S	4	1	2	3
Subject T	4	1	3	2
Subject U	4	2	1	3
Subject V	4	2	3	1
Subject W	4	3	1	2
Subject X	4	3	2	1

Table 5.11: Ordering of structures used to produce lists for each similarity measure that were assessed by study participants.

Structure 1	Structure 2	Structure 3	Structure 4	Structure 5
NSC 2212 <chem>C34H16O2</chem>	NSC 8134 <chem>C10H14O3</chem>	NSC 18084 <chem>C14H19N3O5S</chem>	NSC 90321 <chem>C22H14</chem>	NSC 90799 <chem>C8H17N</chem>

as a Likert scale, was used so that the study participant could communicate their responses in a simple and effective manner. For the evaluation of similarity the value 1 was considered to be “not similar” and the value 7 was considered to be “very similar”. For the correctness of ordering measurement, a value of 1 was considered to be “incorrect” while a value of 7 was considered to be “correct”.

To collect this information a web-based interface was developed using the Jmol applet [63] and FormMail [64]. The Jmol applet provided an interactive Java-based visualization environment for the chemical structures that worked in any web-browser. Using the Jmol applet the query structure and all of the structures that were generated in the “top ten” lists were visualized in 3D for the study participants. Figures 5.7 and 5.8 show screenshots from the web-based interface that was developed for this study.

Before the results could be presented in an organized manner, the raw data obtained from the e-mail messages generated from the interactive web-based interface needed to be formatted and analyzed. This section shows examples of the results that were obtained and provides an overview of the calculations performed. Figure 5.9 shows a sample e-mail message that was generated during the survey when a subject was viewing and scoring one of the lists generated by the standard cosine measure.

Top 10 Search Results


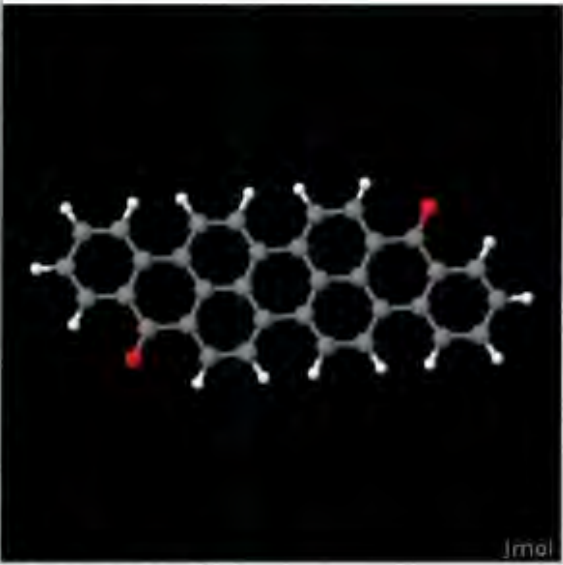
Structure 2204	Structure 5205
 Jmol	 Jmol
<input type="radio"/> 1	<input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7
Not Similar	Similar

Figure 5.7: Sample pairwise similarity scoring web-form as used in the human evaluation with the query structure on the left and the comparison structure on the right. Note that users were able to zoom in and out and rotate each structure if needed through the web-form.

Correctness In List Ordering						
<input checked="" type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7
Incorrect						Correct

Submit

Figure 5.8: Sample correctness of list scoring web-form as used in the human evaluation.

chemicalsurvey@noreply.com () <chemicalsurvey@noreply.com>
Chemical Structure Evaluation Survey - FORMULA
May 14, 2009 10:43:28 AM NDT
Mark Staveley

Below is the result of your feedback form. It was submitted by
(chemicalsurvey@noreply.com) on Thursday, May 14, 2009 at 10:43:28

OUTFILE5 8011 43420 7

OUTFILE5 8011 31533 7

OUTFILE5 8011 22823 6

OUTFILE5 8011 22817 7

OUTFILE5 8011 58740 6

OUTFILE5 8011 58741 5

OUTFILE5 8011 221160 5

OUTFILE5 8011 221626 5

OUTFILE5 8011 199435 5

OUTFILE5 8011 210249 5

OUTFILE5 listscore: 6

mysubmit Submit

Figure 5.9: Sample e-mail generated from the web-based study interface.

E-mails similar to the one in Figure 5.9 were generated for 20 different lists during the course of the survey for each participant. With the 24 participants there were 480 resulting survey e-mail messages that were collected and processed. Each e-mail that was received contained 11 user submitted values, namely the assessed similarity scores for each structure within the “top ten” list and the assessed score for the correctness of the ordering of the list. Appendix B contains all of the user generated data that was obtained from each of the 480 survey e-mail messages.

5.3.6 Data Analysis and Interpretation

After compiling and organizing the raw data, it was important to look at the accuracy of the results that were obtained from the study participants. Within the various “top ten” lists, there were structures that were known to be exactly the same as the query structures. By assessing how well the subject participants scored these values, it became possible to measure the accuracy at which the study participants were assessing similarity. On average, the 24 different study participants assessed the similarity of the structures that were exactly the same as 97.4% ($\sigma = 0.6\%$).

This result demonstrated two things: Firstly it showed that there is a certain error factor here as even experts in the field of chemistry have difficulty in assessing chemical similarity, and secondly it provided additional insight as to the quality of the results that were obtained through the human evaluation. In the statistical component of the evaluation, the similarity measures were assessed on their ability to find structures that matched exactly. The data obtained through the human evaluation

enables the assessment of the different similarity measures with respect to finding similar structures, as relevance judgements can now be defined based on the scores provided by the study participants.

Taking into account the ability of the study participants to identify exact structure matches, a user score threshold of 66% (a value of 4 or higher on the Likert scale that was used throughout the study) was chosen as a basis for assessing similarity. All structures that had an average assessed score of 66% or greater were considered to be similar. Using this threshold it was possible to compare the different similarity measures based on the number of similar structures that were identified within the different “top ten” lists. Table 5.12 shows the number of similar structures that were identified, excluding exact matches, for each of the different test structures and different similarity measures by study participants. Note that the standard cosine measure identifies the greatest number of structures with a user-assessed similarity score of 66% or greater. Exact matches are excluded from the counts as the purpose of this value is to highlight the abilities of the different metrics to go beyond the task of finding an exact match. In the case of all query structures, if there were exact matches within the test collection, then they were identified with a score of 1.0 (as mentioned in subsection 5.2.4).

Throughout the human evaluation, participants were also asked to assess the correctness of the ordering of the different “top ten” lists that were presented to them. Table 5.13 shows the average user-assessed list correctness score, and standard deviation, for each of the lists produced in the human evaluation. It is important to note

Table 5.12: Number of similar structures that were identified by study participants excluding exact matches, for each of the different test structures by the four different similarity measures using a threshold of 66%.

Structure	cosine (Contextual + CATI)	cosine (Standard + CATI)	Tanimoto (CATI)	Tanimoto (chemical fingerprints)
$C_{34}H_{16}O_2$	1	1	0	0
$C_{10}H_{14}O_3$	2	3	3	3
$C_{14}H_{19}N_3O_5S$	0	0	0	0
$C_{22}H_{14}$	0	4	0	0
$C_8H_{17}N$	0	0	0	0
Total	3	7	3	3

that the lists contained only the first ten chemical structures returned, regardless if there were more than ten structures with the same similarity score.

As can be seen, the standard cosine measure consistently has a better user assessed list order correctness score, with a smaller standard deviation, than the scores given to the lists generated using the industry standard Tanimoto measure with chemical fingerprints. In some cases the list order correctness results are based upon lists where the metric has scored all of the items within the list as having perfect similarity scores (test structure 4 with the contextual cosine measure - Figure B.2, test structures 1,2, and 4 with the Tanimoto CATI measure - Figures B.7 and B.8, and test structures 2,4 and 5 with the Tanimoto measure that uses chemical fingerprints - Figures B.10, B.11 and B.12 - note there were no cases of this with the standard cosine measure). An additional example of this can be seen in Figures B.21 and B.16. In this case the first figure (B.21) shows the first five structures returned using the Tanimoto measure that uses chemical fingerprints. All five of the structures were given a score of 1.0 by

Table 5.13: The average user-assessed list correctness score, (and standard deviation), for each of the lists produced in the human evaluation.

Structure	cosine (Contextual + CATB)	cosine (Standard + CATB)	Tanimoto (CATB)	Tanimoto (chemical fingerprints)
$C_{31}H_{16}O_2$	53% (8%)	66% (5%)	36% (11%)	53% (16%)
$C_{10}H_{11}O_3$	78% (7%)	78% (0%)	47% (4%)	60% (8%)
$C_{11}H_{19}N_3O_5S$	48% (3%)	52% (4%)	47% (5%)	45% (9%)
$C_{22}H_{11}$	50% (17%)	73% (2%)	38% (3%)	53% (7%)
$C_8H_{17}N$	44% (9%)	66% (2%)	74% (0%)	40% (6%)
Average List Correctness Score	55%	67%	48%	50%
Average Standard Deviation	8.8%	2.6%	4.6%	9.2%

the metric. This is in contrast to the second figure (B.16) which shows the first five structures returned using the standard cosine measure. Only the first four structures were given a score of 1.0 by the metric. This example also shows trends that were observed with the histograms, namely the granularity in which the structures are differentiated by the similarity measure.

The issue with the lists where all of the scores are 1.0 lies with the assessment of the ordering. In these cases the ordering is based on the order in which they are found within the indices. Even though the value of assessing the order in these cases can be questioned, these results were still presented in this thesis as it highlights performance differences in the metrics that echo back to the precision scores of the different metrics. Should a metric be unable to distinguish structures beyond a certain point, then a deficiency in the metric has been highlighted. These same deficiencies

can also be seen when reviewing the different histograms that show the distribution of the similarity scores of the different metrics when using the test structures from the statistical evaluation.

5.4 Functional Group Investigation

The studies presented thus far concentrate on using various properties, either chemical fingerprints or the CATI topological descriptors, and their ability to assess chemical similarity with a variety of metrics. However, these descriptive properties are not only used as components of metrics to produce similarity scores, but can be also used to identify components and classify chemical structures.

The CATI and CBTI descriptors, as presented in this thesis, are topological descriptors that use computed information to capture different types of chemical information (for example bond information, and the different types of atoms within the structure). In Section 4.2.4, the CATI descriptors were treated as “words” within standard information retrieval measures. Following this analogy, CBTI descriptors can be thought of as “chemical phrases”. This is one approach that can be taken when reviewing functional groups, as functional groups are made up of distinct “chemical words” and “chemical phrases”.

To evaluate the appropriateness of this analogy, different functional groups were represented in terms of CATI and CBTI descriptors to determine if they could be uniquely identified. Table 5.14 summarizes the functional groups that were reviewed

during this analysis and shows what descriptors are required for identification. As an example, it was determined that some functional groups, such as sulfides, could be identified by only using CATI descriptors, whereas other functional groups, such as esters, required additional information that is found within the CBTI descriptors for successful identification.

Table 5.14: Functional group listing and the information required for identification.

Functional Group	CATIs	CBTIs
Alkenes	•	
Alkynes		•
Alkanes		•
Aromatic		•
Alcohol		•
Carboxyl		•
Ester		•
Thioester		•
Ether	•	
Halide	•	
Amine	•	
Nitro		•
Thiol	•	
Sulfide	•	
Nitrile	•	
Aldehyde		•
Ketone	•	
AcylHalide		•
Amide		•
Acid Anhydride		•

This shows a great deal of promise, as one of the traditional methods to identify functional groups was through some type of substructure search involving analysis of molecular subgraphs [65]. By creating indices based on the CATI and CBTI descrip-

tors within chemical structures, a new type of comprehensive functional group searching tool can be created. Because of the proposed architectural framework and the information already found within the Multi-Component Data Representation Scheme, this type of functionality could be implemented with minimal design and coding efforts.

Figure 5.10 through Figure 5.16 provide more detail about the actual CATI and CBTI descriptors needed to identify the different functional groups. As can be seen, the presence of some functional groups can be determined by the occurrence of a single CATI descriptor. In all figures, the items shown in boldface are the key components that are used to identify the functional group with either CATI or CBTI descriptors.

Amine

	Z	C	H
C-N(-H)-H	7	(9,0)	
C-N(-H)-C	7	(19,0)	
C-N(-C)-C	7	(24,0)	

CATI

7(9,0)
7(19,0)
7(24,0)

A structure that contains at least one of the above CATI descriptors contains an amine functional group.

Nitro Compound

	Z	C	H	Z	C	H	Z	C	H
C-N(-O)=O	7	(34,0)		8	(5,-1)		8	(5,-1)	

CBTI

8(5,-1)-7(34,0) x 2

A structure that contains two 8(5,-1)-7(34,0) CBTI descriptors which have a common 7(34,0) CATI descriptor has a nitro compound functional group.

Thiol

	Z	C	H	Z	C	H
C-SH	16	(7,0)		1	(14,0)	

CATI

16(7,0)

A structure that contains at least one 16(7,0) CATI descriptor has a thiol functional group.

Figure 5.10: CATI and CBTI descriptors used to identify the presence of amines, nitro compounds, and thiols within chemical structures.

Sulfide

	$Z \begin{smallmatrix} \text{S} \\ \text{S} \end{smallmatrix}$
C-SC	16(12,0)

CATI

16(12,0)

A structure that contains at least one 16(12,0) CATI descriptor has a sulfide functional group.

Nitrile

	$Z \begin{smallmatrix} \text{N} \\ \text{N} \end{smallmatrix}$	$Z \begin{smallmatrix} \text{N} \\ \text{N} \end{smallmatrix}$
C-C#N	6(14,-2)	7(4,0)
H-C#N	6(10,-2)	7(4,0)

CATI

6(14,-2)

6(10,-2)

A structure that contains either a 6(14,-2) or 6(10,-2) CATI descriptor has a nitrile functional group.

Aldehyde

	$Z \begin{smallmatrix} \text{O} \\ \text{O} \end{smallmatrix}$	$Z \begin{smallmatrix} \text{O} \\ \text{O} \end{smallmatrix}$	$Z \begin{smallmatrix} \text{O} \\ \text{O} \end{smallmatrix}$
C-C(=O)-H	6(25,-1)	8(4,-1)	1(4,0)

CATI

6(25,-1)

A structure that contains at least one 6(25,-1) CATI descriptor has an aldehyde functional group.

Figure 5.11: CATI and CBTI descriptors used to identify the presence sulfides, nitriles, and aldehydes within chemical structures.

Carboxyl

	Z	z	z
C-C(-OH)=O	6(34,-1)	8(7,0)	8(4,-1)

CBTI

8(7,0)-6(34,-1)
8(4,-1)-6(34,-1)

A structure that contains an 8(4,-1)-6(34,-1) CBTI descriptor, and an 8(7,0)-6(34,-1) CBTI descriptor where both CBTI descriptors share a common 6(34,-1) CATI descriptor contains a carboxyl functional group.

Ester

	Z	z	z
C-C(-OC)=O	6(34,-1)	8(12,0)	8(4,-1)

CBTI

8(12,0)-6(34,-1)
8(4,-1)-6(34,-1)

A structure that contains an 8(4,-1)-6(34,-1) CBTI descriptor, and an 8(12,0)-6(34,-1) CBTI descriptor where both CBTI descriptors share a common 6(34,-1) CATI descriptor contains an ester functional group.

Thioester

	Z	z	z
C-C(-SC)=O	6(58,-1)	16(12,0)	8(4,-1)

CBTI

16(12,0)-6(58,-1)
8(4,-1)-6(58,-1)

A structure that contains an 16(4,-1)-6(58,-1) CBTI descriptor, and an 16(12,0)-6(58,-1) CBTI descriptor where both CBTI descriptors share a common 6(58,-1) CATI descriptor contains a functional group that is a sulfur variant of an ester.

Figure 5.12: CATI and CBTI descriptors used to identify the presence of carboxyl groups, esters, and thioesters within chemical structures.

Ether

	Z	C	E
C-O-C	8	(12,0)	

CATI

8(12,0)

A structure that is not defined as an ester, and still has a 8(12,0) CATI descriptor is identified as a structure with an ether functional group.

Halide

	Z	C	E
C-F	9	(4,0)	
C-Cl	17	(4,0)	
C-Br	35	(4,0)	
C-I	53	(4,0)	
C-At	85	(4,0)	

CATI

9(4,0)
17(4,0)
35(4,0)
53(4,0)
85(4,0)

A structure that contains at least one of the above CATI descriptors contains a halide functional group.

Figure 5.13: CATI and CBTI descriptors used to identify the presence of ethers and halides within chemical structures.

Alkenes

	Z	ζ	ξ	Z	ζ	ξ
H ₂ C=CH ₂	6(9,-1)			6(9,-1)		
HCC=CH ₂	6(19,-1)			6(9,-1)		
C ₂ C=CH ₂	6(24,-1)			6(9,-1)		
C ₂ C=CCH	6(24,-1)			6(19,-1)		
C ₂ C=CC ₂	6(24,-1)			6(24,-1)		

CBTI

6(9,-1)-6(9,-1)
 6(19,-1)-6(9,-1)
 6(24,-1)-6(9,-1)
 6(24,-1)-6(19,-1)
 6(24,-1)-6(24,-1)

A structure that is made up of only carbon and hydrogen atoms which contains at least one of the above CBTI descriptors is an alkene.

Alkynes

	Z	ζ	ξ	Z	ζ	ξ
HC#CH	6(7,-2)			6(7,-2)		
CC#CH	6(12,-2)			6(7,-2)		
CC#CC	6(12,-2)			6(12,-2)		

CBTI

6(7,-2)-6(7,-2)
 6(12,-2)-6(7,-2)
 6(12,-2)-6(12,-2)

A structure that is made up of only carbon and hydrogen atoms which contains at least one of the above CBTI descriptors is an alkyne.

Alkanes

	Z	ζ	ξ	Z	ζ	ξ
C-C(-C)(-C)-C	6(40,0)					
C-C(-C)(-C)-H	6(35,0)			1(4,0)		
C-C(-C)(-H)-H	6(25,0)			1(4,0)		
C-C(-H)(-H)-H	6(10,0)			1(4,0)		
H-C(-H)(-H)-H	6(-10,0)			1(4,0)		

CATI

6(40,0)
 6(35,0)
 6(25,0)
 6(10,0)
 6(-10,0)
 1(4,0)

A structure that is made up of only the above CATI descriptors is an alkane.

Figure 5.14: CATI and CBTI descriptors used to identify structures that are alkenes, alkynes or alkanes.

Aromatic**C₆H₆ Ring****CBTI**

$Z = 6$
 $\zeta = -1$
 $\zeta = 19$
 $6 \times 6(19,-1)$ and part of the same ring

$6(19,-1)-6(19,-1)$
 $6(>19,-1)-6(19,-1)$

Alcohol

	$Z \quad \zeta \quad \eta$	$Z \quad \zeta \quad \eta$
H-(H-)C-O-H	8(7,0)	6(33,0)
C-(H-)C-O-H	8(7,0)	6(43,0)
C-(C-)C-O-H	8(7,0)	6(48,0)

CBTI

$8(7,0)-6(33,0)$
 $8(7,0)-6(43,0)$
 $8(7,0)-6(48,0)$

A structure that contains at least one of the above CBTI descriptors. The 8(7,0) CATI descriptor that is part of the above CBTI descriptors cannot be bonded to a 6(34,-1) or 6(29,-1) CATI descriptor. If these conditions are met, then the structure contains an alcohol functional group.

Figure 5.15: CATI and CBTI descriptors used to identify structures that are either aromatic (e.g. C₆H₆) or that contain alcohols.

Ketone

	Z	C	E
C-C(=O)-C	6(30,-1)	8(4,-1)	

CATI

6(30,-1)

A structure that contains at least one 6(30,-1) CATI descriptor has a ketone functional group.

Acyl Halide

	Z	C	E
C-C(=O)-F	9(4,0)	8(4,-1)	6(37,-1)
C-C(=O)-Cl	17(4,0)	8(4,-1)	6(61,-1)
C-C(=O)-Br	35(4,0)	8(4,-1)	6(115,-1)
C-C(=O)-I	53(4,0)	8(4,-1)	6(169,-1)
C-C(=O)-At	85(4,0)	8(4,-1)	6(265,-1)

CBTI

9(4,0)-6(37,-1)
17(4,0)-6(61,-1)
35(4,0)-6(115,-1)
53(4,0)-6(169,-1)
85(4,0)-6(265,-1)

The presence of 1 or more of these CBTI descriptors identifies the presence of an acid halide functional group

Amide

	Z	C	E
H-C(=O)-N(-H)-H	6(27,-1)	8(4,-1)	7(9,0)
H-C(=O)-N(-H)-C	6(27,-1)	8(4,-1)	7(19,0)
H-C(=O)-N(-C)-H	6(27,-1)	8(4,-1)	7(19,0)
H-C(=O)-N(-C)-C	6(27,-1)	8(4,-1)	7(24,0)
C-C(=O)-N(-H)-H	6(32,-1)	8(4,-1)	7(9,0)
C-C(=O)-N(-H)-C	6(32,-1)	8(4,-1)	7(19,0)
C-C(=O)-N(-C)-H	6(32,-1)	8(4,-1)	7(19,0)
C-C(=O)-N(-C)-C	6(32,-1)	8(4,-1)	7(24,0)

CBTI

8(4,1)-6(32,-1)
7(9,0)-6(32,-1)
7(19,0)-6(32,-1)
7(24,0)-6(32,-1)

The presence of a 8(4,1)-6(32,-1) CBTI descriptor, that shares a 6(32,-1) CATI descriptor with either a 7(9,0)-6(32,-1) or 7(19,0)-6(32,-1) or 7(24,0)-6(32,-1) CBTI descriptor identifies the presence of an amide functional group.

Acid Anhydride

	Z	C	E
O=C(-C)-O-C(-C)=O	6(34,-1)	8(4,-1)	8(12,0)

CBTI

8(12,0)-6(34,-1)
8(4,-1)-6(34,-1)

The presence of two 8(12,0)-6(34,-1) CBTI descriptors that are connected to the same 8(12,0) CATI descriptor, as well as the presence of two 8(4,-1)-6(34,-1) CBTI descriptors identifies the presence of an acid anhydride functional group

Figure 5.16: CATI and CBTI descriptors used to identify the presence of ketones, acyl halides, amides, or acid anhydrides within a chemical structure.

5.5 Discussion

Upon examining the results obtained from both the statistical evaluation (precision and recall data) and the human evaluation (assessment of pairwise similarity vs. list rank, and the correctness in list ordering), it is observed that the null hypothesis is rejected. The first result that demonstrates the performance difference of the two metrics is the precision data that is observed when using the 19 test structures in the statistical evaluation. If there was no difference in the metrics, then the precision data would be the same for the Chem-DRSM metrics and the Tanimoto metric that uses chemical fingerprints. Furthermore, the histogram data demonstrates that the metrics are not just linear shifts or translations of one another, re-affirming that the metrics are not the same.

The statistical results demonstrate that the metrics are not the same, and if the performance of the metrics was only based on precision then the standard cosine measure that is part of the Chem-DRSM system would have the best performance. However, the statistical evaluation primarily looks at exact matches (as there were no other ways available to independently confirm relevance judgements) and the metrics are not just boolean queries that look for an exact structure. The human evaluation results extend the results obtained throughout the statistical evaluation by consulting with study participants that have expert knowledge. Throughout the study, the human participants identified the standard cosine metric as their preferred measure (through anecdotal comments), comments which were subsequently reaffirmed when reviewing the results that were obtained through the study. Throughout the study,

the standard cosine measure identified the highest number of structures to be given an average user-assessed similarity score of 66% or better. This is an important statistic as it shows that the standard cosine measure is useful beyond searching for exact matches.

The high quality of the results produced by the standard cosine measure which uses the CATI descriptors can be attributed to two features of this particular measure. The first feature is that the cosine measure (as implemented in the Chem-DRSM system, Section 4.2.4) uses information about the statistical distribution of components when assessing similarity. This is in contrast to the Tanimoto measure that does not use any kind of statistical distribution information. The second feature, is the use of the CATI descriptor. The CATI descriptor (as mentioned in Section 4.1.1) is derived from the topological information and valency of each atom within a given structure. This provides an increased chemical vocabulary that can be used to describe chemical structures. The combination of both of these features has resulted in high quality results, as assessed not only by statistical values (precision and recall) but also by human subjects that are considered to have expert knowledge in the field of chemistry.

Looking forward, the experimentally determined threshold values for the various computationally derived descriptors and the work being done with the identification of functional groups highlights even more ways in which the various descriptors found within the Chem-DRSM system can be extended and applied to activities related to the searching, browsing, and organization of chemical structure information.

Chapter 6

Prototype Comprehensive Computational Chemistry Database

This chapter extends the work that has been completed through the creation of the Chem-DRSM system and presents additional work that builds on the core functionality of the tools that are found within the Chem-DRSM system. This chapter is divided into a number of sections, namely data representation, integration, and enhanced chemical information classification.

This work is still in the very early stages, but the promising results that have been observed warrant that it be included with the work contained within this thesis. In particular, it highlights the architectural flexibility and modularity that is inherent with the data-representation scheme at the core of the Chem-DRSM system. It also highlights the range of information that is captured within that data-representation

scheme and illustrates some of the alternative ways that it can be used to enable comprehensive access to chemical information.

6.1 Data Representation

As mentioned throughout this thesis, one of the main challenges faced by digital librarians and managers of digital archives is the automatic processing and classification of data. The Multi-Component Data Representation Scheme found within the Chem-DRSM system provides a universal framework for data management that can be extended very easily to be used with large-scale resources. Furthermore, the modules required to build the different data components that make up this data representation scheme are independent of one another and do not require any specialized computational resources. The presence of these design features mean that researchers and scientists could contribute results very easily, and automatically, to large scale resources (for example a central database) using the tools found within the Chem-DRSM system.

6.2 Integration

By building on the foundation established by the Chem-DRSM system, a fully interactive and dynamic database could be easily constructed and integrated into the computational resources used by chemical researchers. Figure 6.1 illustrates the architecture of such a system. Within the system, there are four key layers (the interface layer, the management layer, the processing layer, and the storage layer) that make up the overall architecture. One key feature of the layered design is that communi-

cation only occurs between adjacent layers. This means that the components within the layers can be updated, modified and added to with minimal implications to the system as a whole.

The only components that are required outside of the Chem-DRSM system to implement a large scale chemical information resource are an appropriately designed interface that supports interactive use and batch processing and a software component that would manage all the communications and interactions between the different components. By including these two modules, a comprehensive computational chemistry database could be deployed.

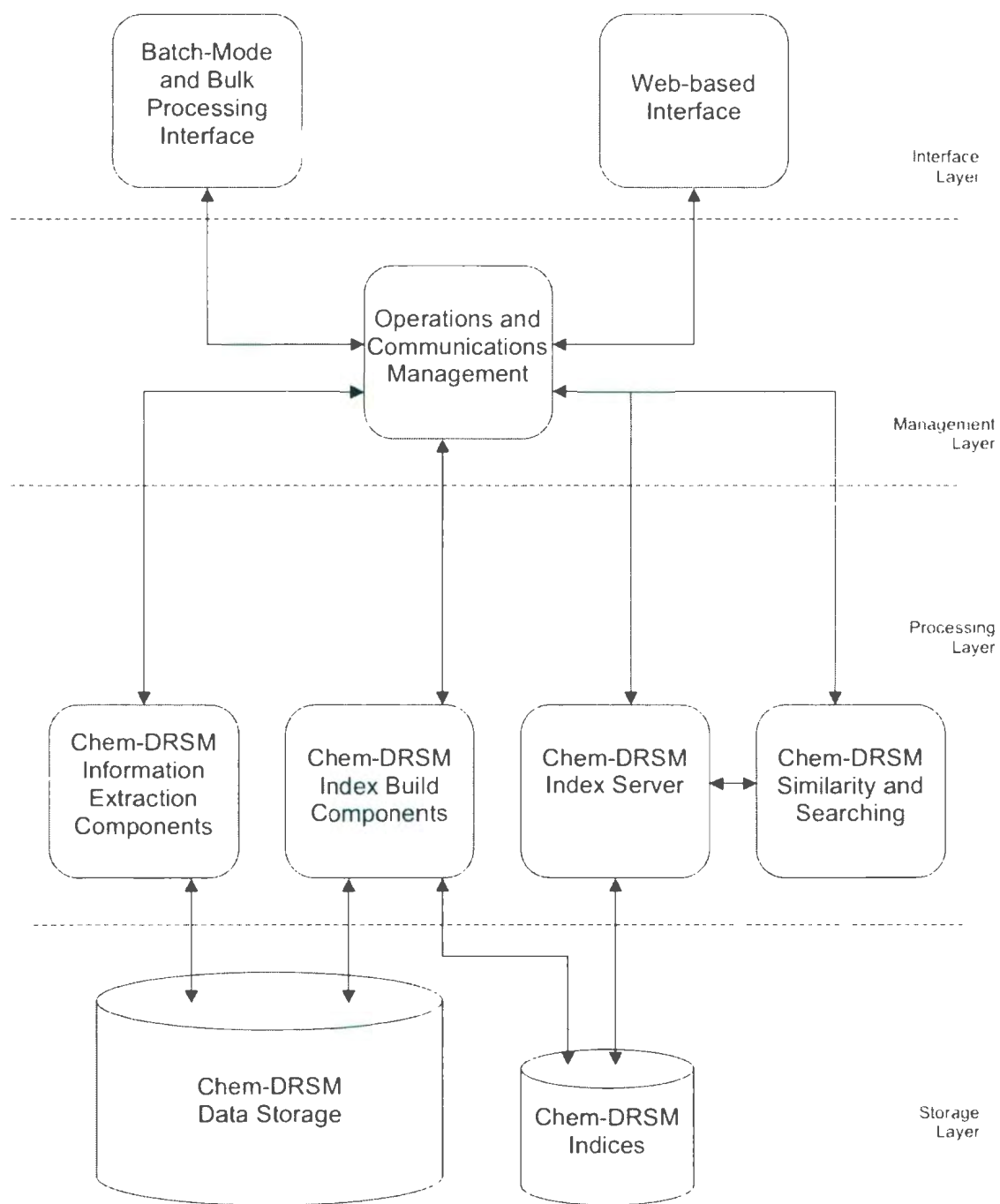


Figure 6.1: Comprehensive Computational Chemistry Database architecture, including Chem-DRSM components.

6.3 Enhanced Chemical Information Classification

Once the basic functionality of a comprehensive computational chemistry database has been implemented, work can be done to improve how the chemical information within the database is classified and used. Initial investigations done as part of this thesis have shown significant promise in extending the use of the components found within the Multi-Component Data Representation Scheme to allow for comprehensive searching tools that could identify functional groups within chemical structures and also dynamically cluster and link the chemical structures within the database in different ways. This section specifically discusses how the similarity scores that are derived from the CATI descriptors can be used to create clusters of chemical structures and establish links between them.

Having a system that can automatically cluster and classify structures found within a database based on user customizable properties could be particularly useful for automating the process of identifying lead compounds. If a chemist is required to test all structures that have a certain property or functional group, it might make more sense to test very diverse samples from the collection of available structures as opposed to testing a set of structures that are all similar in composition. This type of selection can assist a chemist's productivity as they may only be able to test a small number of compounds at a time (for example, having 5000 test structures and only being able to test 80 at a time – see case study in Subsection 6.3.1). By allowing the system to cluster the structures based on diversity, the chemist, in this particular example, could increase their productivity by improving the efficiency of each experimental run.

One possible way of creating clusters of chemical structures is to use the pairwise similarity scores of all structures within the database to generate links. This is consistent with methods already established in the area of Hypertext and link-creation methodology [66]. As an example, a pairwise similarity score of 0.9 could be used as the link threshold. This means that a relationship is defined between two structures when their similarity is assessed to be greater than or equal to 0.9.

Since it is possible for a relationship to be only in one direction (where $A = B$, but $B \neq A$), there are two different types of relationships to be considered. The first is an "outbound" relationship, where the similarity score between "A" and some other structure is greater than 0.9. The second type of relationship, an "inbound" relationship, where the similarity score between some other structure and "A" is greater than 0.9. Based on this link definition, the number of inbound and outbound relationships can be determined for each structure.

Once the inbound and outbound relationships are determined then the structures can be assessed by reviewing differences in the number of incoming and outgoing relationships. The difference in the number of inbound and outbound relationships, (calculated using $|In - Out|$), provides some insight into how significant the structure is within the entire collection of structures. Structures that have equal number of inbound and outbound links are not considered to be as important as those that have a disproportionate number of either inbound or outbound links. This difference in the number of relationships can be used to provide a guide as to what structures should

be considered or not considered for various clusters or summary lists.

Figures 6.2 and 6.3 show two different structures ($C_6H_5N_2O_2Cl$ and $C_{10}H_{11}N_2O_3Cl$) that when compared using the contextual cosine measure, produce different similarity results depending on how they are compared (A to B, or B to A). When comparing structure A to structure B, the similarity is scored as 0.93 (or 93%), and when comparing B to A the similarity is scored as 0.90 (or 90%). It is differences like this that can impact how links are established since, depending on the threshold used, the structures may or may not have links between them or they may only have one-way links. Subsection 6.3.1 outlines work that was done in collaboration with a researcher from Merck Frosst to create clusters and links using the contextual cosine measure. This work was done on a collection of 5000 carboxylic acids that are supplied by different vendors and the Chem-DRSM system was used to find the 80 most diverse structures out of the given 5000.

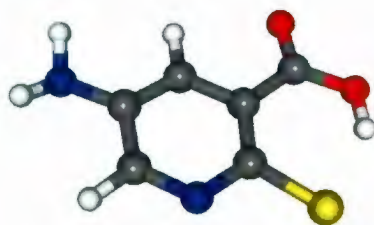


Figure 6.2: Example Structure A - $C_6H_5N_2O_2Cl$.



Figure 6.3: Example Structure B - $C_{10}H_{11}N_2O_3Cl$.

6.3.1 Practical Example of Clustering Groups of Chemical Structures - Carboxylic Acids

Large collections of chemical structure data can be an obstacle for many experimental chemists. In particular, the process of determining what structures should be considered as candidates for an experiment can, in some cases, be a difficult decision. This subsection discusses how the Chem-DRSM system can assist chemists with this decision-making process by automatically determining the most interesting (or representative) structures from a given list of candidates. It is worth noting that this project would not have happened without the assistance of the Combinatorial Chemistry Gordon Research Conference (GRC). It was at this conference where the author and Dr. Rejean Fortin (a Senior Researcher with Merck Frosst Canada Ltd.) met and were able to talk about research within an environment that supported and encouraged collaboration. Dr. Fortin provided a list of 5000 commercially available

carboxylic acids for analysis.

For the steps outlined below, all timings provided are from a computer with a 2.4 GHz dual core processor and 2GB of RAM (unless otherwise specified). It is also important to note that these steps do not have to be repeated every time work is done with a particular set of structures. Only the process relating to “Relationship Determination” (6.3.1.6) would need to be repeated with every subsequent iteration.

6.3.1.1 SMILES to SDF to XYZ

Upon initial review of the list, it was determined that no three-dimensional Cartesian coordinate information was present. This was troublesome, as all of the tools and metrics within the Chem-DRSM system use the information contained within Cartesian coordinate files. Upon further investigation, a novel method was discovered that allowed for the conversion of a SMILES to three-dimensional Cartesian coordinates. Experimental open source software was used to perform this conversion. The first program, **smi2sdf**, generates a rough set of three-dimensional Cartesian coordinates, contained within a SDF data file, using an iterative refinement procedure. The coordinates are then optimized using a MMFF94 force field by the second program **engine**. These two programs have been grouped together under the name **smi23D** [67] and the whole process has been built upon a methodology described by Ballester and Graham Richards [68].

In-house code was written to automatically extract, process, and convert the SMILES representation for each structure from the original list into Cartesian coordinates

(SMILES to SDF to Cartesian coordinates). This exercise highlights the flexibility of the Chem-DRSM system and how different data formats can be easily converted and integrated to work with the system (typically it had been standard practice to have the geometry of the structure in either .XYZ or .SDF files). To translate all 5000 SMILES into Cartesian coordinate data files took approximately 42 minutes to complete.

There were 99 structures out of 5000 (2% of structures) that could not be converted using this method. The result from this process was a SDF data file that contained the three-dimensional coordinates for 4901 structures. This SDF file was subsequently converted to an XYZ data file using Open Babel, and the conversion process took approximately 2 minutes to complete.

6.3.1.2 XYZ file Preparation

As mentioned above, Open Babel has the ability to convert SDF data files to XYZ files. However, the SDF that is produced from the previous processing step is a single file that contains all of the information for all of the 4901 structures. As such, the resulting conversion from Open Babel is a single XYZ data file that contains all of the information for each of the 4901 structures. The information within the XYZ file can then be processed by Chem-DRSM and the appropriate descriptors obtained. A decision was made to convert the large XYZ data file into smaller XYZ files, where an XYZ file was created for each chemical structure. The reason for creating the individual XYZ files for each structure was so that the structures can be independently read and processed. The Chem-DRSM system took approximately 5

minutes to complete this conversion / translation process.

6.3.1.3 Computational Chemistry Calculations

Using the Chem-DRSM input files produced in the previous step, all of the structures were subsequently processed in batch by customized Perl scripts that executed the various Chem-DRSM commands and stored the results from the different calculations into suitably named output files. During the processing and analysis of each structure, CATI descriptors were identified and recorded. The entire process took approximately 40 minutes to complete when using a computer with a 2.8GHz processor and 4GB of RAM. It is important to mention that this part of the process can be completed in parallel (where linear speedup has been observed). As an example of the parallelism performance, this same work could be completed in 10 minutes when using a machine with 4 processors.

During this process only 4632 structures produced valid results. The 269 structures that did not produce valid results had errors resulting from charge / multiplicity errors. Unless charge information is specifically mentioned (SMILES does not contain charge information) the structure's charge was assumed to be 0. The resulting collection of 4632 structures is 93% of the size of the original collection of 5000 structures. The processing rate is typically 120 structures / minute / processor.

6.3.1.4 Index Creation

Indices were then created for the 4632 different structures based upon the frequency and types of CATI descriptors that were found within the structures. This index cre-

ation process took approximately 3 minutes to complete and the observed processing rate was typically 2000 structures / minute.

6.3.1.5 Similarity Scoring

As context is important when establishing clusters and links, the contextual cosine measure as implemented with the Chem-DRSM system, was used to compute the contextual similarity between each structural pair. In this case, it took approximately 72 hours to complete all of the similarity calculations at a rate of 90 scores / second on a single processor.

As with the computational chemistry calculation step (Section 6.3.1.3), this process is designed to support a parallel implementation with linear scaling. The result from this process was a 4632 x 4632 matrix where all the entries in the matrix corresponded to the similarity scores for that relationship (i.e row 1, column 2 is the relationship where structure 1 is compared to structure 2 and row 2, column 1 is the relationship where structure 2 is compared to structure 1). For the carboxylic acid collection (4632 structures), the average similarity score of all of the pairwise comparisons is 0.53, and the processing time is estimated at 5400 comparisons / minute / processor.

6.3.1.6 Relationship Determination

Using the contextual Cosine measure and methods already established with Hypertext and link-creation methodology, pathways connecting the different structures were identified. For our purposes, a score of 0.9 was used as a the link threshold. This meant that a relationship was defined when there was a similarity score greater than

0.9 between two structures. There are two different types of relationships to be considered, the first is an "outbound" relationship. This is where the similarity score between "A" and some other structure is greater than 0.9. The second type of relationship, an "inbound" relationship, is where the similarity score between some other structure and "A" is greater than 0.9. Based on this link definition, the number of inbound and outbound relationships were determined for each structure. The analysis of the quantities of inbound and outbound links provides insight as to what structures should be considered within the summary of the carboxylic acid collection.

6.3.1.7 Discussion

Initial feedback of this work was provided by Dr. Fortin and a number of key points were raised. First, the concept of having the same structure identified as being a "member" or present in more than one set of structures was something that researchers at Merck Frosst working with Dr. Fortin had not considered. This is in contrast to our method, which allows multiple links to be established between structures subsequently giving each structure a different browsing or linking pathway. Second, the refinement of computationally derived descriptors (such as nuclear repulsion energy and origin-invariant nuclear second-moment) was encouraged so that additional information about particular features within a structure could be evaluated. It was highlighted that in some cases, the overall shape or energy of a structure might be misleading, particularly when chemists are only interested in a particular functional group or combination of atoms. Overall however, the study itself was viewed as promising and further work in the area of automatically clustering, linking, and summarizing collections of chemical structures was encouraged.

6.4 Summary

Although this chapter is a brief overview of some of the design considerations, functional enhancements, and properties found within the data used and generated by the Chem-DRSM system, it shows that the foundation laid by the Chem-DRSM system can be easily extended to further support chemical research in a way that is aligned with the comprehensive needs of chemical researchers. This in itself was very encouraging, as the work done with Merck Frosst was completed before development work on the Chem-DRSM system was completed.

Chapter 7

Discussion and Future Work

In order to better organize the conclusions that are made, this chapter has been divided into three sections. The first section discusses conclusions that can be drawn from the results of the precision-recall evaluation of the similarity measures, the data obtained through the evaluation of the distribution of the similarity scores, and the human evaluation of the similarity measures, as shown in Chapter 5. The second section discusses system performance issues with the Chem-DRSM system, and the third section discusses further experiments that could be conducted, proposes suggestions for design enhancements and the refinement of the Chem-DRSM system, and presents an overview of what is being conceptually called a National Comprehensive Computational Chemistry DataBase (NCCC'DB).

7.1 Conclusions Drawn from Experimental Results

For each similarity measure (the contextual cosine measure, the standard cosine measure and the Tanimoto measure all using the CATI descriptor and the Tanimoto mea-

sure with chemical fingerprints), three methods were used for evaluation: precision-recall data (Section 5.2.4), distribution data relating to similarity scores and different query structures (Section 5.2.5), and a human evaluation (Section 5.3) where ranked lists produced by each of the similarity measures were evaluated in terms of the similarity of each item in the list to the query structure and the correctness of list ordering. Five different structures were used throughout the human evaluation component, and the statistical evaluation was conducted using 19 different structures.

7.1.1 Precision-Recall Statistical Evaluation of Performance

In order to determine the precision and recall values for a given chemical structure, relevance judgements are required. In the case of the statistical evaluation, the relevance judgements were based on the structures that were exactly the same (equivalent canonical SMILES and equivalent InChIs).

In terms of recall, all four similarity measures achieved 100% recall by assigning a score of 1.0 to all the structures that were exactly the same. However, the precision of the four different similarity measures was different. The measure with the highest average precision, across the 19 structures, was the standard cosine measure with an average precision score of 92% (standard deviation of 17%). This is in contrast to the average precision observed across the 19 structures by the Tanimoto measure that used chemical fingerprints, where the average precision score was 75% (standard deviation of 31%).

Although exact matches can be found using either InChI or canonical SMILES de-

scriptors, precision scores provides measurable insight into the how accurate the different similarity measures can be. Too many false positives will serve to dilute the quality of the results presented to users when they perform searching and browsing activities. The impact of which can be seen in the analysis of the distributions of similarity scores and the human evaluation.

7.1.2 Analysis of the Distribution of Similarity Scores

The precision and recall data was generated by assessing the performance of the different similarity measures in their ability to find exact matches. This data, although important, only provided a partial picture of the quality of the results produced by the different similarity measures. By analyzing the distribution of the similarity scores produced by the different measures, it becomes possible to further assess how granular the similarity measures are in assessing similarity and identifying distinguishing properties.

Upon reviewing the similarity scores generated for the 19 query structures as they compared to the structures within the test collection, it was observed that the standard cosine measure had a tendency to be more granular in nature when determining how similar two structures are, as compared to the Tanimoto measure that makes use of chemical fingerprints. Similar behaviour was observed when reviewing results produced by the Tanimoto measure which uses the CATI descriptors. The fact that both of these measures exhibit similar behaviour leads one to conclude that this behaviour is primarily attributed to the nature of the Tanimoto measure where only the presence / absence of a feature is considered, and not its quantity or statisti-

cal significance; a conclusion that re-iterates the importance of taking into account the statistical distribution and weighting of the components being used to determine similarity.

7.1.3 Results and Observations of Performance from the Human Evaluation

Upon completion of the precision and recall statistical evaluation, it was observed that the standard cosine measure which uses the CATI descriptors had the highest average precision across the 19 structures and the smallest variance when compared to the contextual cosine measure, the Tanimoto measure which uses the CATI descriptors and the Tanimoto measure that uses Chemical Fingerprints. However, a further evaluation was required, as even though the results could be considered correct in terms of relevance judgments and precision-recall data, there was still the issue of the quality of the results that did not score 100% similarity and the ordering of the results being returned by the similarity measures.

Participants in the human evaluation of these measures were asked to provide two different types of assessments. The first type of assessment provided by the study participants scored the similarity between the search structure and each of the structures within a list of the first ten structures returned by the measure (structures are returned by the metric in descending order of similarity score). The second type of assessment provided by the study participants scored the ordering ability of the similarity measure.

According to the results recorded by study participants, the standard cosine measure which uses the CATI descriptor found and placed more similar structures in the first ten structures returned by the measure. Furthermore, the quality of list ordering provided by the standard cosine measure had a smaller standard deviation amongst the study participants, and for four out of five of the test structures the standard cosine measure had the highest quality of list ordering, as scored by study participants. For the case where the standard cosine measure did not have the highest quality list ordering, it had the second highest list ordering score as compared to the other three similarity measures.

Although limited, both of these indicators from the human evaluation component along with the precision and recall information from the statistical evaluation component, and the distributions of similarity scores demonstrates that, when considering our test cases, the standard cosine measure produces ranked lists of candidate structures that are more appropriate.

7.2 System Performance Differences

Although the work being presented within this thesis has concentrated primarily on the results obtained from the four different studies (computational descriptor thresholds, statistical evaluation of metrics, human evaluation of metrics, and functional group searching assessment) there are many other factors to be considered. This section describes some of the issues that influence the creation and storage of the

information required by the different similarity measures.

7.2.0.1 Storage

The first factor to consider is disk storage space. Indices are required by the two cosine and the CATI Tanimoto similarity measures. Furthermore, the type of measure being used can have an influence on the size of the indices that are required to be stored, not only on disk but also in main memory during the calculation process. Table 7.1 shows the sizes of the different indices that are used by each of the similarity measures that are part of the Chem-DRSM system. A value for the Tanimoto chemical fingerprint measure is also shown. This value has been approximated by creating the same type of indices that are required by the CATI Tanimoto measure, except with 128-bit chemical fingerprints as determined by OpenBabel.

Table 7.1: Differences in Storage Requirements for Different Indexing Schemes when Indexing 178,175 Structures.

Index type	CATI (cosine)	CATI (Tanimoto)	Chemical Fingerprints (Tanimoto)
Index size	59 MBytes (uncompressed)	18 MBytes (uncompressed)	6.1 MBytes (uncompressed)
	12.7 MBytes (compressed)	4.4 MBytes (compressed)	1.7 MBytes (compressed)
Required space per structure	347 Bytes (uncompressed)	106 Bytes (uncompressed)	36 Bytes (uncompressed)
	75 Bytes (compressed)	26 Bytes (compressed)	10 Bytes (compressed)

The results in Table 7.1 indicate a significant difference in the sizes of the indices required by the different similarity measures. Even when compression is used, the

ratios in size still remain constant, with the Tanimoto chemical fingerprint measure using the least amount of storage space and the cosine measures requiring the most index space. This is to be expected since the CATI descriptors are more complex than chemical fingerprints and there is additional information about statistical distribution of the CATI descriptors that is required.

Index space is an important consideration, as computing performance suffers when the required indices cannot be stored entirely in main memory during processing. The size of the indices required by the two Tanimoto measures are smaller than those required by the cosine measures, but the cosine indices are all still quite small when considering the baseline memory size that is currently available on most commodity compute servers. Table 7.1 also provides an approximate calculation (number of structures divided by the size of the indices) that shows the size required for each chemical structure within the indices. Even with the larger disk space requirements of the cosine measure, it is still less space than the average size of a data file representing the Cartesian coordinates for a chemical structure (approximately 2048 Bytes).

7.2.0.2 Index Creation

Not only should the size of the indices be considered when designing such a system, but processing times should be considered also. As the number of terms in the index increases, so does the construction time. This was taken into consideration throughout the construction of the Chem-DRSM system and is one of the reasons for its modular design. Not only can the build process be completed without impacting the similarity measures, but the indices can be constructed in parallel thereby further

reducing the time required to complete the build process. This is important as data should be easily accessed and integrated into such a system.

7.2.0.3 Similarity Measure Calculations

Another type of performance issue to be considered relates to the computation of similarity scores. The Tanimoto measure is a computationally fast measure as there is no statistical weighting being considered and only the terms present within the structures being compared are considered. Although this makes for faster query completion times and smaller indices, it does not do the best job in measuring chemical similarity.

When considering the two cosine measures there are two things that influence the time associated with the calculation of similarity values. First, there is the number of descriptors associated with each chemical structure. This not only influences the index size, which in turn influences the memory requirements, but it also influences the number of mathematical operations required to produce the similarity score. Even a small difference in the number of terms, in this case contextual vs. standard, can influence the number of summations required in the cosine calculation. Second, the nature of the similarity measure being used also influences computational time requirements. As already observed and discussed, the Tanimoto measure has fewer calculations that need to be completed, as compared to the cosine measures, before a similarity score can be determined. The cosine measure has already demonstrated itself to be a very thorough yet computationally intensive measure of document similarity in the area of Information Retrieval [1].

However, there are measures that approximate different values within the cosine measure and the resulting approximate cosine measure has the ability to decrease the number of calculations that are required, which in turn decreases the computation time. One example of such an approximate cosine measure is the work done by de Kretser et al [69]. In de Kretser's work, the components of the weighting value that require any data pertaining to the frequency of a descriptor with respect to the entire document collection is approximated using a logarithmic value. This type of approximation has two major benefits. First, the reduction of computation time to determine the similarity of a query and a document. Second, less maintenance is required when building collections and collection indices as the process of determining how many structures have a given property (a process that would need to be redone every time new data is added) would no longer need to be completed. Although this type of cosine measure approximation work has been done within the context of English language and textually based information retrieval, it has not been applied to chemical information retrieval. This area is of interest for future work as the measures used in the evaluation were just simple adaptations of the cosine measure that used standard weighting schemes with CATI descriptors. Special CATI-based measures need to be further developed in order to take advantage of the unique properties and characteristics that are found in chemical structures, and the information, both topological and computational, that is contained within them.

One possibility for extending the accuracy of the CATI based measures could involve the combination of topological information and computational information to form one single measure instead of having to have a topological term-based measure

that is further refined using computationally derived information. Another area for consideration is to use the presence and absence of general topological features, such as the number of rings, to improve similarity assessments.

7.3 Future Work

The results obtained from the evaluation of the Chem-DRSM system are encouraging. Although there are areas where the results can be extended and refined, the results nonetheless show a great deal of promise with the work discussed thus far. This section presents three different avenues for future work involving the Chem-DRSM. The first subsection comments further on the human evaluation and discusses possible extensions of this work. The second subsection outlines areas where the initial investigation of the Chem-DRSM system could be extended, and the third subsection discusses areas for the modification and extension of the Chem-DRSM system, along the proposition of a National Computational Chemistry DataBase (NCC'CDDB) that could be used to assist computational and chemical researchers on a very large scale.

7.3.1 Extending the Human Evaluation

Although the results relating to the Human Evaluation are quite encouraging, there is more work that can be done to build on the results obtained thus far. Firstly, it is important to point out the demographics of the subjects participating in the study. Even though there were chemists from many different disciplines, there was a noticeable lack of medicinal chemists and biologists from the subject population. It is important to consider extending the Human Evaluation to include medicinal chemists

and biologists as this type of functionality offered in the Chem-DRSM system would be very useful for them.

Furthermore, it is worthwhile to consider further studies involving the assessment of the ranking algorithms. One option may be to give subjects a unranked list of structures and ask them to rank the structures themselves, then comparing the resulting list to those produced by the different similarity measures.

7.3.2 Further Experiments involving the Chem-DRSM system

There are a number of areas where refinements and extensions to the investigation conducted in this thesis would be appropriate. The first area relates to the test collection of structures that are used for testing and evaluating information retrieval performance. The most difficult part about the precision-recall evaluation was the process of determining appropriate relevance judgements. The relevance judgements used could be considered restrictive as the judgements were made by two descriptors that are designed for finding exact structural matches (InChI and canonical SMILES) instead of being assessed in a standard way by experts in the field of chemistry. It would be useful to undertake the process of designing a purpose built test corpus for testing these different chemical similarity measures, something that could possibly follow a similar model to the TREC (Text Retrieval Conference) initiative [70].

Another area that warrants further investigation is the notion of establishing clusters

and links within collections of chemical structure collections based on the results from the different similarity measures. As presented in Section 6.3, there are very definite real-world applications that could benefit from the assistance provided by systems that assist with the organization and clustering of groups of chemical structures. Implementing experiments to test and evaluate the clustering abilities of the different chemically based similarity measures would certainly be of interest and relevance to the work presented in this thesis.

Related to experiments that extend the results from chemically based similarity measures would be additional experiments to determine if there are appropriate approximations that could be made within the various measures, as discussed in Section 7.2.0.3 and if any correlation could be determined between the resulting similarity scores and the activity of a given chemical structure.

A preliminary investigation was conducted to examine the correlation between AIDS activity and similarity scores produced by the standard cosine measure with the CATI descriptors and the similarity scores produced by the Tanimoto measure that uses chemical fingerprints. The National Cancer Institute has AIDS activity data for $\sim 28,000$ of the chemical structures that are found within the test collection that has been used throughout the evaluation of the Chem-DRSM system. These 28,000 structures are classified either as active, moderately active or inactive.

As a performance indicator for the initial study, results from a separate study by Martin et al [71] were used. According to the work done by Martin, there is only a

30% chance that structures with a similarity score of 85% or greater, as assessed by the Tanimoto measure with chemical fingerprints to an active structure, will themselves be active. As such, the threshold of 85% or greater was used in this initial investigation.

An initial analysis of the $\sim 28,000$ structures used for this preliminary investigation revealed that only 185 of the $\sim 28,000$ structures were classified as active. When using the Tanimoto measure with chemical fingerprints none of the active structures within the test collection were assigned a score of 0.85 or greater when they were compared to the query structure (randomly chosen out of the 185 active structures). This is in contrast to the standard cosine measure with the CATI descriptors which assigned a score of 0.85 or greater to 15 of the active structures. This represented a precision of 24% and a recall of 8%. Even though additional work needs to be completed to fully and quantitatively evaluate the relationship between the different similarity scores and the biological activity of a chemical structure, this initial result shows promise that the standard cosine measure with the CATI descriptors may be able to successfully identify candidate structures that have a similar biological activity.

7.3.3 Future Development and Applications

There are a number of potential applications for the Chem-DRSM system. The initial integration of the Chem-DRSM system into a prototype comprehensive computational chemistry database, as shown in Chapter 6, has identified a number of areas where enhancements and refinements of the current Chem-DRSM system could be made. Similarly, this prototype comprehensive computational chemistry database has high-

lighted areas where this type of system, if deployed on a large-scale, could be of great benefit to a large number of researchers.

This subsection has two parts: the first part discusses possible enhancements and design changes to the Chem-DRSM system, and the second part discusses strategies and suggestions as to how the prototype system in Chapter 6 could be extended to a large-scale resource.

7.3.3.1 Chem-DRSM Version 2.0

There are a number of areas where the performance and design of the Chem-DRSM system could be improved. However, these performance enhancements and design changes were not realized and deemed feasible until after the initial prototype system had been tested and evaluated. The first area for improvement involves how the indices are created, maintained and stored. The current architecture of the Chem-DRSM system employs a very simple index module. All the indices are created in the form of tab-delimited text files. Tab-delimited files are useful since not only are they easy to read and process, but they are easily imported into relational databases. Similarly, queries from relational databases can also be returned in tab-delimited form. Based on the experimental and performance results observed with the Chem-DRSM system thus far, a logical progression in the evolution of the system would be to store the built indices in a relational database that has the same logical design as what has been implemented with the indices in the flat-file format.

Another area to be considered is the integration of the build process and the user

search process with the schedulers that are used on large scale computer clusters. By integrating the job schedulers during the build process and the information extraction phase, both processes can become even more automated than they are now. Also by dynamically integrating with the job scheduler, the job scheduler would be able to determine the optimal method for the processing of the information instead of having a fixed algorithm that limits how the work can be distributed.

The final two revisions relate to the user-experience with the Chem-DRSM. Consultations with users will need to take place to ensure that appropriate web-based interfaces are designed to support the functionality found within Chem-DRSM system. Currently, interaction with the Chem-DRSM system is done through a textually based shell interface and script files. This can be easily extended through the use of a dynamic language such as Python or through the use of PHP code to integrate with a suitable web-based interface. Consultations with both technical designers and chemists will be required to ensure maximum usability. Also, the same script files and shell interfaces can be extended to support an automated batch interface that can be used for the importation of large data collections and the automated importation of chemical information as it becomes available.

7.3.3.2 National Comprehensive Computational Chemistry DataBase (NCCDB)

The work presented throughout this thesis highlights the strengths of the Chem-DRSM system and how it can be used to support chemical research. As mentioned in the introduction of this thesis, it is envisioned that a nation-wide, and eventu-

ally worldwide, resource of this kind will not only be a source of information for researchers, but it would also be an intelligent system that could automatically collect and classify both public domain and user-contributed data.

By providing a solid foundation for data management, namely the Multi-Component Data Representation scheme, that is combined with intelligent searching and browsing tools, such as the similarity measures within the Chem-DRSM system, the goal of establishing a highly reliable chemical resource that extends beyond the realm of just experimental or patent data is now one step closer to becoming a reality. This thesis has demonstrated various applications of the tools within the Chem-DRSM system, and it has also shown that there are cases where the Chem-DRSM system yields results that are as good as or better than the results (as assessed both statistically and by test subjects with expert level chemistry knowledge) yielded by the Tanimoto measure with chemical fingerprints (an industry standard).

References

- [1] I.H.Witten, A.Moffat, and T.C.Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, first edition, 1994.
- [2] Protein Data Bank. <http://www.rcsb.org/pdb/home/home.do>.
- [3] National Cancer Institute (NCI) database. cactus.nci.nih.gov/ncidb2/.
- [4] National Institutes of Health (NIH). PubChem. <http://pubchem.ncbi.nlm.nih.gov/>.
- [5] ZINC: A Free Database for Virtual Screening. <http://zinc.docking.org/>.
- [6] I.H.Witten, Z.Bray, M.Mahoui, and W.J.Teahan. Text Mining: A New Frontier for Lossless Compression. In *Proceedings Data Compression Conference '99*, IEEE Press, Los Alamitos, CA, pages 198–207, 1999.
- [7] E.Hovy and C.Y.Lin. Automated Text Summarization in SUMMARIST. In *Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization, Topic Identification, Summarization*, pages 18–24, 1997.
- [8] I.H.Witten, G.W.Paynter, E.Frank, G.Gutwin, and C.G.Nevill-Manning. KEA:

- Practical Automatic Keyphrase Extraction. In *Proceedings DL'99, Berkeley CA*, pages 254–255, 1999.
- [9] D.Weininger. SMILES. A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* 28:31–36, 1989.
- [10] D.Weininger, A.Weininger, and J.L.Weininger. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* 29:97–101, 1989.
- [11] A.McNaught. The IUPAC International Chemical Identifier (InChI). *Chemistry International, IUPAC*, 2006.
- [12] H.Bunke. On a Relation Between Graph Edit Distance and Maximum Common Subgraph. *Pattern Recognition Letters*, 18(8):689–694, 1997.
- [13] J.S.Duca and A.J.Hopfinger. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. *J. Chem. Inf. Comput. Sci.* 41:1367–1387, 2001.
- [14] R.Guha, M.T.Howard, G.R.Hutchison, P.Murray-Rust, H.Rzepa, C.Steinbeck, J.K.Wegner, and E.L.Willighagen. The Blue Obelisk Interoperability in Chemical Informatics. *J. Chem. Inf. Mod.*, 46:991–998, 2006.
- [15] The Open Babel Package version 2.2.0. <http://openbabel.sourceforge.net/>.
- [16] M.Feher and J.M.Schmidt. Property Distributions: Differences between Drugs,

- Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* 13:218–227, 2003.
- [17] H.Lowell, L.H.Hall, and L.B.Kier. Issues in Representation of Molecular Structure: The Development of Molecular Connectivity. *J. Mol. Graphics Modelling*, 20:4–18, 2001.
- [18] M.Randic. The Connectivity Index, 25 Years After. *J. Mol. Graphics Modelling*, 20:19–35, 2001.
- [19] H.Weiner. Structural Determination of Paraffin Boiling Point. *The Journal of the American Chemical Society*, 69:17–20, 1947.
- [20] M.Randic. On the Characterization of Molecular Branching. *The Journal of the American Chemical Society*, 97:6609–6615, 1975.
- [21] L.B.Kier and L.H.Hall. *Molecular Connectivity in Structure-Activity Analysis*. Wiley, New York, 1986.
- [22] C.Hansch and A.Leo. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. American Chemical Society, 1995.
- [23] D.J.Livingstone. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* 40:195–209, 2000.
- [24] L.H.Hall and L.B.Kier. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd (Editors), 2:367–422, 1991.

- [25] L.H.Hall and L.B.Kier. The Electrotological State: An Atom Index for QSAR. *Quantitative Structure-Activity Relationships*, 10:43–51, 1991.
- [26] Chemical Abstracts Service (CAS) Registry and CAS Registry Numbers. <http://www.cas.org/expertise/cascontent/registry/regsys.html>.
- [27] Molecular Design Limited (MDL) Information Systems - symyx technologies. <http://www.symyx.com>.
- [28] A.Dalby, J.G.Nourse, W.D.Hounshell, A.K.I.Gushurst, and D.L.Grier et al. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* 32:244–255, 1992.
- [29] R.B.King. *Applications of Graph Theory and Topology in Inorganic Cluster and Coordination Chemistry*. CRC Press - Taylor & Francis Group, first edition, 1992.
- [30] N.Trinajstić. *Chemical Graph Theory*. CRC Press - Taylor & Francis Group, second edition, 1992.
- [31] C.Jochum and J.Gasteiger. Canonical Numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* 17(2):113–117, 1977.
- [32] O.Ivanciuc. Canonical Numbering and Constitutional Symmetry. *Handbook of Chemoinformatics*. J. Gasteiger (Editor), pages 139–160, 2003.
- [33] R.Grossman, P.Kasturi, D.Hamelberg, and B.Liu. Experimental Studies of the Universal Chemical Key (UCK) Algorithm on the NCI Database of Chemical

- Compounds. In *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, page 244, 2003.
- [34] P.Murray-Rust and H.Rzepa. Chemical Markup, XML and the World Wide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* 39:923–942, 1999.
- [35] C.A.James, D.Weininger, and D.Delany. Daylight Theory Manual - Daylight Chemical Information Systems Inc. Dec 2008. <http://www.daylight.com/dayhtml/doc/theory/theory.toc.html>.
- [36] C.A.James, D.Weininger, and D.Delany. Daylight SMARTS Example. Daylight Chemical Information Systems Inc. http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html.
- [37] SMILEStm Toolkit. Daylight Chemical Information Systems Inc. http://www.daylight.com/products/smiles_kit.html.
- [38] M.Stevens, A.Lenstra, and B.deWeger. Vulnerability of Software Integrity and Code Signing Applications to Chosen-Prefix Collisions for MD5. Nov 2007. <http://www.win.tue.nl/hashclash/SoftIntCodeSign/>.
- [39] Unofficial InChI FAQ. <http://wwwmm.ch.cam.ac.uk/inchifaq>.
- [40] M.Karelson, V.S.Lobanov, and A.R.Katritzky. Quantum Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* 96:1027–1044, 1996.
- [41] Google Search Engine. <http://www.Google.com>.
- [42] Bing Decision Engine. <http://www.bing.com>.

- [43] New Zealand Digital Library. <http://www.NZDL.org>.
- [44] M.Karelson, V.S.Lobanov, and A.R.Katritzky. Quantum chemical descriptors in qsar/qspr studies. *Chem.Rev.*, 96 (3):1027–1044, 1996.
- [45] S. Wolfram. Wolfram Alpha - A Computational Knowledge Engine, June 2009. <http://www.wolframalpha.com>.
- [46] P.Willett, J.Barnard, and M.Downs. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, 38:983–996, 1998.
- [47] P.Willett and V.Winterman. A Comparison of Some Measures for the Determination of Intermolecular Structural Similarity. *Quant. Struct-Act. Relat.*, 5:18–25, 1986.
- [48] T.T.Tanimoto. IBM Internal Report, Nov 1957.
- [49] G.M.Downs and P.Willett. Similarity Searching in Databases of Chemical Structures. *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, (Editors), 7:1–66, 1995.
- [50] M.F.Lynch and P.Willett. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.*, 18(3):154–159, 1978.
- [51] J.W.Raymond, E.J.Gardiner, and P.Willett. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.*, 42:305–316, 2002.

- [52] X.Girones, D.Robert, and R.Carbo-Dorca. TGSA: A Molecular Superposition Program Based on Topo-Geometrical Considerations., *J. Comput. Chem.*, 22:255–263, 2001.
- [53] L.Spialter. The Atom Connectivity Matrix (ACM) and its Characteristic Polynomial (ACMCP). *J. Chem. Doc.*, 4(4):261–269, 1964.
- [54] I. Mayer. Charge, Bond Order and Valence in the Ab Initio SCF Theory. *Chem. Phys. Letters*, 97(3):270–274, 1983.
- [55] A. Szabo and N.S.Ostlund. *Modern Quantum Chemistry*. Dover Publishing, 1996.
- [56] W.Hehre, R.F.Stewart, and J.A.Pople. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *J. Chem. Phys.*, 51:(6):2657–2664, 1969.
- [57] J.W.Hollett, A.Kelly, and R.A.Poirier. Quantum Mechanical Size and Steric Hindrance. *J. Phys Chem.*, 110:13884–13888, 2006.
- [58] J.B.Hendrickson, P.Huang, and A.G.Toczko. Molecular Complexity - A Simplified Formula Adapted to Individual Atoms. *J. Chem. Inf. Comput. Sci.*, 27:63–67, 1987.
- [59] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R.

Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision B.05, 2003. Gaussian, Inc., Pittsburgh PA.

- [60] R. A. Poirier and J. W. Hollett. MUNgauss (Fortran 90 version), 2007. Chemistry Department, Memorial University of Newfoundland, St. John's, NL, A1B 3X7. With contributions from S. D. Bungay, A. El-Sherbiny, T. Gosse, D. Keefe, A. Kelly, C. C. Pye, D. Reid, K. Saputantri, M. Shaw, M. S. Staveley, Y. Wang and J. Xidos.
- [61] D. R. Flower. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.*, 38(3):379–386, 1998.
- [62] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly Analysis of a Very Large Topically Categorized Web Query Log. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 321–328, New York, NY, USA, 2004. ACM.

- [63] Jmol: An Open-Source Java Viewer for Chemical Structures in 3D.
<http://www.jmol.org/>.
- [64] M.Wright. FormMail Version 1.91. <http://www.scriptarchive.com/>.
- [65] J.M.Barnard. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* 33(4):532–538, 1993.
- [66] M. Massimo. An Evaluation of Automatically Constructed Hypertexts for Information Retrieval. *Inf. Retr.*, 1(1-2):91–114, 1999.
- [67] K. Gilbert and R. Guha. Smi23D, a Conversion Utility for Converting SMILES to 3D. <http://rguha.net/res.html>.
- [68] J. Ballester P., J. and W. G. Graham Richards. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* 28(10):1711–1723, 2007.
- [69] O.deKretser, A.Moffat, T.Shimmin, and J.Zobel. Methodologies for Distributed Information Retrieval. In *Proceedings of the 18th International Conference on Distributed Computing Systems, Amsterdam, NL*, pages 66–73, 1998.
- [70] Text REtrieval Conference (TREC). <http://trec.nist.gov>.
- [71] Y. C. Martin, J. L. Kofron, and L. M. Traphagen. Do Structurally Similar Molecules have Similar Biological Activity. *J. Med. Chem.* 45:4350–4358, 2002.

Appendix A

Statistical Data

The histogram data for the distribution of the query result scores when using the statistical evaluation test structures with the different similarity measures is shown in Tables A.1 to A.19. The discussion pertaining to these results can be found in Chapter 5, Section 5.2.

Table A.1: Distribution of similarity scores produced by different similarity measures when structure NSC 131564 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.4)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.4)
1.00	6	6	12	12
0.95	0	6	286	0
0.90	0	16	190	0
0.85	13	22	586	8
0.80	0	64	376	4
0.75	141	107	162	11
0.70	49	109	108	18
0.65	490	144	356	28
0.60	1527	122	1552	66
0.55	2126	139	2126	104
0.50	7520	571	21336	582
0.45	6027	2789	62246	800
0.40	15879	3585	32330	1598
0.35	21660	4329	20662	3984
0.30	42445	6171	26830	9484
0.25	64419	9505	7890	31198
0.20	53428	12857	6910	61052
0.15	29978	15714	56966	79836
0.10	11142	19482	60	89052
0.05	1957	25328	46588	55826
0.00	1264	77109	68778	22684

Table A.2: Distribution of similarity scores produced by different similarity measures when structure NSC134422 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.4)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.4)
1.00	6	8	31	8
0.95	0	6	1700	0
0.90	0	11	8419	0
0.85	0	93	17906	0
0.80	0	176	4997	14
0.75	6	337	5489	0
0.70	0	571	1084	0
0.65	11	911	895	81
0.60	25	1163	528	2
0.55	0	1495	301	171
0.50	203	1968	90	480
0.45	0	2695	23209	0
0.40	684	3279	9496	2405
0.35	56	4089	1646	2245
0.30	1918	5269	232	8231
0.25	10114	6482	17733	16143
0.20	20867	8779	1861	23991
0.15	41651	12202	66	25011
0.10	109267	17060	3	35198
0.05	65667	22576	0	55626
0.00	9596	88975	79489	8569

Table A.3: Distribution of similarity scores produced by different similarity measures when structure NSC 134438 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	9	9	9	9
0.95	0		1	0
0.90	0	7	34	0
0.85	0	33	121	0
0.80	2	90	837	1
0.75	25	171	1776	14
0.70	32	220	920	40
0.65	129	293	459	119
0.60	455	386	281	868
0.55	767	379	165	1192
0.50	3559	471	401	4186
0.45	4043	652	1130	3485
0.40	11534	879	2323	9756
0.35	19689	1441	5381	12636
0.30	29012	2312	13207	15316
0.25	50990	4909	28067	21916
0.20	54085	9316	22614	25689
0.15	43293	14367	28360	31722
0.10	33192	22943	21947	30079
0.05	7847	35006	29776	11683
0.00	1408	84291	20366	9464

Table A.4: Distribution of similarity scores produced by different similarity measures when structure NSC' 152324 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.4)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.4)
1.00	15	15	57	15
0.95	0	6	3654	0
0.90	0	33	2302	0
0.85	9	60	574	0
0.80	18	146	222	23
0.75	17	213	26296	0
0.70	144	318	3195	0
0.65	71	520	787	137
0.60	457	783	13740	4
0.55	741	1237	11096	351
0.50	1224	1494	6473	466
0.45	688	1929	912	0
0.40	2651	2697	144	1852
0.35	2690	3565	43	1801
0.30	5797	5058	15	6836
0.25	13467	7243	12	11024
0.20	29219	10291	1	15029
0.15	67202	11960	0	15980
0.10	107281	12978	0	30749
0.05	26519	9660	0	84143
0.00	1864	107969	108652	9768

Table A.5: Distribution of similarity scores produced by different similarity measures when structure NSC 153096 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	5	3	4	8
0.95	0	18	337	0
0.90	4	31	790	8
0.85	4	46	472	11
0.80	0	112	2675	37
0.75	9	169	2719	56
0.70	23	293	4531	134
0.65	63	465	3142	242
0.60	263	673	1878	639
0.55	692	858	4801	141
0.50	2938	1030	6683	3121
0.45	3639	1655	2392	3110
0.40	10528	2437	2802	8430
0.35	18650	3094	20448	15047
0.30	29875	4367	39389	21463
0.25	50465	8308	41432	38774
0.20	48510	17177	22721	34262
0.15	40389	28194	2191	23096
0.10	32800	33892	5875	14637
0.05	18030	34775	7912	5462
0.00	3184	10578	4981	9194

Table A.6: Distribution of similarity scores produced by different similarity measures when structure NSC 167530 is the query.

Interval	cosine (Contextual + CATB) Chem-DRSM Equation (4.7)	cosine (Standard + CATB) Chem-DRSM Equation (4.6)	Tanimoto (CATB) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	26	5	13729	5
0.95	0	5	36683	0
0.90	0	39	11637	0
0.85	0	112	6014	0
0.80	0	246	2400	0
0.75	0	423	1860	0
0.70	0	773	1719	0
0.65	0	1023	1283	131
0.60	0	1460	1017	0
0.55	0	1938	696	0
0.50	327	2605	565	385
0.45	0	3722	443	0
0.40	0	4893	430	1309
0.35	0	6751	483	0
0.30	1165	9508	185	2699
0.25	2773	12797	71	14572
0.20	5035	16427	13	28557
0.15	8381	21056	0	48247
0.10	65194	23299	39723	52009
0.05	96668	17949	0	21449
0.00	80502	53144	59224	8812

Table A.7: Distribution of similarity scores produced by different similarity measures when structure NSC 4765 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	19	9	13	9
0.95	0	4	173	0
0.90	0	20	426	0
0.85	35	49	825	17
0.80	68	79	1773	0
0.75	106	143	902	97
0.70	320	173	196	30
0.65	325	223	136	378
0.60	1059	341	44	852
0.55	1213	381	6589	362
0.50	5664	693	7316	3841
0.45	5417	1042	5885	2973
0.40	11382	1489	11524	5601
0.35	18086	2502	9021	6911
0.30	34111	3475	13644	11967
0.25	51655	4773	8599	20247
0.20	50749	6985	2979	27373
0.15	43648	11320	824	37440
0.10	28826	15697	37657	37501
0.05	5961	20395	50732	14430
0.00	1427	108382	18917	8146

Table A.8: Distribution of similarity scores produced by different similarity measures when structure NSC 169899 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.4)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.4)
1.00	6	6	6	6
0.95	0	2	9	0
0.90	11	5	33	1
0.85	0	11	28	4
0.80	43	32	229	2
0.75	78	84	1147	12
0.70	61	229	1140	34
0.65	313	347	1179	78
0.60	697	544	5808	233
0.55	906	1024	15389	376
0.50	3573	1874	10024	1051
0.45	3573	2781	4127	1180
0.40	9401	4120	7142	3577
0.35	16712	5513	14674	6977
0.30	28145	8255	35184	12306
0.25	56726	12313	33616	24614
0.20	63083	17873	11962	34141
0.15	50402	23848	11046	41737
0.10	21787	30017	12132	30234
0.05	4017	33993	8156	10111
0.00	537	35304	5144	11501

Table A.9: Distribution of similarity scores produced by different similarity measures when structure NSC 170347 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (1.7)	cosine (Standard + CATI) Chem-DRSM Equation (1.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	13	6	12	14
0.95	0	4	184	0
0.90	4	41	234	0
0.85	0	41	24	26
0.80	15	18	6	0
0.75	16	20	214	48
0.70	8	46	3292	20
0.65	34	184	5538	184
0.60	109	308	5098	174
0.55	218	487	884	570
0.50	1301	709	398	2924
0.45	1480	812	106	3848
0.40	5772	964	22	7826
0.35	12672	1132	16	15126
0.30	27595	1103	30044	23650
0.25	69599	1257	31238	45076
0.20	64501	2375	17836	61186
0.15	42687	6285	42280	74140
0.10	26299	12942	43740	73394
0.05	6411	29844	130810	29272
0.00	1337	119597	44374	18572

Table A.10: Distribution of similarity scores produced by different similarity measures when structure NSC 209826 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (1.7)	cosine (Standard + CATI) Chem-DRSM Equation (1.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	8	8	8	8
0.95	0	1	16	0
0.90	0	1	315	5
0.85	0	47	936	0
0.80	8	107	2528	12
0.75	8	285	4395	73
0.70	0	123	5653	69
0.65	40	677	6717	172
0.60	81	1103	13709	501
0.55	150	1732	15509	716
0.50	789	2593	11959	1937
0.45	938	3695	17027	1713
0.40	3267	5054	7814	1994
0.35	8630	7017	22295	8246
0.30	16364	9504	18704	12930
0.25	41329	12650	1482	26239
0.20	68639	16056	11063	30439
0.15	70264	20187	117	31378
0.10	38831	22959	0	26695
0.05	9243	26052	10877	21455
0.00	1482	47718	24051	10593

Table A.11: Distribution of similarity scores produced by different similarity measures when structure NSC' 210746 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	9	9	9	9
0.95	0	1	11	0
0.90	7	13	73	9
0.85	0	22	375	3
0.80	36	172	847	32
0.75	74	255	1840	114
0.70	112	358	4481	215
0.65	251	673	9285	320
0.60	1032	1346	12185	830
0.55	2013	2191	14803	1638
0.50	8488	3536	18586	3491
0.45	9722	5311	15083	4621
0.40	26161	7290	13798	10410
0.35	32329	10296	17415	17578
0.30	36673	13406	11609	26093
0.25	49707	16688	11354	30716
0.20	42311	18999	11268	29183
0.15	28538	21979	9882	21353
0.10	17338	23171	9846	14290
0.05	4323	20265	5578	7728
0.00	917	32194	9844	9542

Table A.12: Distribution of similarity scores produced by different similarity measures when structure NSC 15309 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (1.7)	cosine (Standard + CATI) Chem-DRSM Equation (1.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	20	6	13	175
0.95	0	18	1803	0
0.90	65	39	5697	0
0.85	0	72	7891	332
0.80	178	193	9756	157
0.75	289	295	9744	829
0.70	261	534	11231	257
0.65	650	871	13684	1509
0.60	1508	1434	18970	2643
0.55	1883	2004	13844	2703
0.50	5550	2762	11843	8291
0.45	4523	3912	11485	6584
0.40	11147	5106	10008	13393
0.35	17771	7010	13380	17949
0.30	29023	9554	9691	25442
0.25	53582	12841	5519	32770
0.20	53211	17584	6289	24525
0.15	43544	22450	218	16600
0.10	28648	27447	4	9833
0.05	6789	29753	9005	5448
0.00	1429	34290	5100	8735

Table A.13: Distribution of similarity scores produced by different similarity measures when structure NSC 1880 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (1.7)	cosine (Standard + CATI) Chem-DRSM Equation (1.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	8	6	7	10
0.95	0	6	85	0
0.90	9	35	465	0
0.85	14	54	993	17
0.80	31	138	2119	49
0.75	53	205	4755	52
0.70	158	406	7732	179
0.65	151	811	14184	135
0.60	576	1237	11469	598
0.55	603	1821	7703	729
0.50	3405	2676	16319	3472
0.45	3265	3795	8123	2902
0.40	8289	5049	12401	5726
0.35	15430	6581	16427	9118
0.30	28895	8773	7603	13682
0.25	54982	11049	1459	23603
0.20	63834	13749	42549	26740
0.15	50411	16962	2271	23965
0.10	24568	21234	448	40025
0.05	4420	29446	16353	18022
0.00	969	54142	4110	9151

Table A.14: Distribution of similarity scores produced by different similarity measures when structure NSC 525079 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	3	3	3	4
0.95	0	3	116	0
0.90	6	27	1240	46
0.85	15	17	2299	0
0.80	58	169	4492	114
0.75	151	316	8587	406
0.70	378	602	10121	497
0.65	452	970	11784	1301
0.60	1147	1698	13943	3280
0.55	1897	2843	14033	4369
0.50	6798	4230	12747	12655
0.45	8547	5391	13816	9273
0.40	22297	7220	18517	20424
0.35	30508	9721	18134	22476
0.30	37944	12503	19444	21929
0.25	49439	15688	11680	26598
0.20	39487	19112	2914	18977
0.15	31527	22203	4389	14178
0.10	23171	24100	559	9343
0.05	5174	23956	5413	4453
0.00	1072	27370	3944	7852

Table A.15: Distribution of similarity scores produced by different similarity measures when structure NSC 623441 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	22	3	26	77
0.95	0	37	4470	0
0.90	98	104	3825	0
0.85	0	224	7479	311
0.80	280	427	7748	0
0.75	234	773	13172	875
0.70	594	1024	9216	1208
0.65	1323	1473	12196	343
0.60	2557	2078	48561	2212
0.55	2594	2818	18020	2166
0.50	10342	3686	7337	6352
0.45	8664	4634	7680	5047
0.40	20108	6009	6037	9157
0.35	29776	7747	1672	12571
0.30	37810	9925	1170	17991
0.25	50460	13001	409	29254
0.20	41940	16859	2049	30325
0.15	32383	21117	8601	24513
0.10	15512	25488	1478	16982
0.05	4371	23176	12302	10816
0.00	1003	37572	1727	7975

Table A.16: Distribution of similarity scores produced by different similarity measures when structure NSC 26613 is the query.

Interval	cosine (Contextual + CxTf) Chem-DRSM Equation (4.7)	cosine (Standard + CxTf) Chem-DRSM Equation (4.6)	Tanimoto (CxTf) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	6	5	5	59
0.95	0	102	584	0
0.90	28	250	1366	296
0.85	0	468	3239	0
0.80	91	997	5416	819
0.75	242	1513	9846	661
0.70	213	1810	9729	1187
0.65	599	2115	11322	1171
0.60	1647	2451	10998	3533
0.55	2006	2979	12900	3386
0.50	6076	3620	12568	10276
0.45	5705	4368	8878	7059
0.40	16137	5249	11319	15001
0.35	24370	6710	20070	17812
0.30	34395	8984	26880	19830
0.25	55461	11830	9394	27223
0.20	53132	15297	5104	23810
0.15	33239	19758	5189	18752
0.10	21071	25314	3194	13522
0.05	4638	30231	6006	5355
0.00	1015	34124	4168	8123

Table A.17: Distribution of similarity scores produced by different similarity measures when structure NSC 79367 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	13	7	7	7
0.95	0	0	11	0
0.90	1	13	59	0
0.85	0	11	25	8
0.80	15	17	222	1
0.75	16	83	629	28
0.70	8	80	1147	30
0.65	34	211	735	69
0.60	109	331	2000	415
0.55	218	544	8126	131
0.50	1301	900	12146	2102
0.45	1480	1441	9178	2214
0.40	5772	2158	14344	1238
0.35	12672	3148	7131	7535
0.30	27595	5018	8136	11874
0.25	69599	7050	8629	21136
0.20	64501	9273	1305	26748
0.15	42687	11509	25837	35065
0.10	26299	15298	17877	40577
0.05	6411	26860	36564	16277
0.00	1337	94220	21067	9720

Table A.18: Distribution of similarity scores produced by different similarity measures when structure NSC 8134 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	11	5	5	18
0.95	0	11	95	0
0.90	10	52	725	35
0.85	0	62	1044	0
0.80	43	208	4431	153
0.75	1	335	3659	276
0.70	174	530	3310	366
0.65	57	946	11853	851
0.60	447	1494	7155	2076
0.55	345	1938	10236	2287
0.50	2046	2755	15900	7036
0.45	1870	3635	9540	5472
0.40	4102	4991	6228	11502
0.35	6569	6280	1545	15509
0.30	13318	7782	9717	18515
0.25	36634	10137	19828	28666
0.20	61005	13382	40665	26316
0.15	76258	17592	12238	24020
0.10	47801	23040	6374	18393
0.05	8176	34093	9274	7870
0.00	1204	48907	4353	8814

Table A.19: Distribution of similarity scores produced by different similarity measures when structure NSC 90799 is the query.

Interval	cosine (Contextual + CATI) Chem-DRSM Equation (4.7)	cosine (Standard + CATI) Chem-DRSM Equation (4.6)	Tanimoto (CATI) Chem-DRSM Equation (3.1)	Tanimoto (chemical fingerprints) OpenBabel Equation (3.1)
1.00	22	3	5	8
0.95	0	58	5517	0
0.90	0	462	6557	0
0.85	0	1263	25842	0
0.80	0	1642	8986	43
0.75	117	1698	5901	0
0.70	0	1838	2199	99
0.65	11	2082	5696	18
0.60	383	2393	2633	232
0.55	0	2708	2406	509
0.50	1198	3310	692	928
0.45	0	3963	7483	895
0.40	2936	4543	12488	2818
0.35	2259	5245	2988	1649
0.30	9999	6291	637	10520
0.25	23963	7600	380	19849
0.20	37846	9376	20544	28068
0.15	51309	11858	3210	33159
0.10	84594	15512	14452	29915
0.05	38690	22218	44355	37646
0.00	6744	74112	5204	8819

Appendix B

Human Evaluation Data

The data that was obtained from the responses given by participants of the human evaluation study (Section 5.3) is shown in Table B.1 and Figures B.1 to B.22. Personal information that could identify study participants was not collected and the study subjects were assigned an ID number based on the order in which they completed the various tasks (for example, ID 1234 corresponds to contextual cosine measure first, standard cosine measure second, Tanimoto CATI measure third, and Tanimoto Chemical Fingerprint measure fourth). Table B.1 summarizes the demographic data of study participants and also shows their ID numbers. Following Table B.1 is all of the data collected for each of the 20 “top ten” lists, data which in turn was used to assess the four different similarity measures (Figures B.1 to B.12). Additionally, Figures B.13 to B.22 show images of the first five structures that were returned for each of the five test structures using the standard cosine measure (Chem-DRSM system) and the Tanimoto measure with chemical fingerprints (OpenBabel).

Table B.1: ID number, gender, education and area of specialization of human study participants.

ID	Gender	Education	Area of Specialization
1234	M	PhD (Chemistry)	Theoretical
1243	M	PhD (Chemistry)	Theoretical / Computational
1324	F	PhD (Chemistry)	Physical
1342	F	MSc (Chemistry)	Theoretical / Computational
1423	M	PhD (Chemistry)	Physical
1432	M	MSc (Chemistry)	Organic / Experimental
2134	F	PhD (Chemistry)	Physical
2143	M	BSc (Chemistry)	Organic
2314	F	PhD (Chemistry)	Theoretical
2341	M	MSc (Computational Science)	Physics / Condensed Matter
2413	M	PhD (Chemistry)	Theoretical / Computational
2431	M	PhD (Chemistry)	Physical / Computational
3124	M	PhD (Chemistry)	Theoretical / Computational
3142	M	MSc (Chemistry)	Physical Chemistry
3214	M	PhD (Chemistry)	Analytical
3241	M	PhD (Chemistry)	Physical / Computational Chemistry
3412	F	PhD (Chemistry)	Crystallography / Inorganic
3421	F	PhD (Chemistry)	Analytical
4123	M	PhD (Chemistry)	Inorganic
4132	M	PhD (Chemistry)	Organic / Experimental
4213	M	PhD (Chemistry)	Organic / Experimental
4231	M	PhD (Chemistry)	Organic
4312	M	PhD (Chemistry)	Theoretical / Computational Chemistry
4321	F	PhD (Chemistry)	Theoretical / Computational Chemistry

Context	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (7)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
Cosine Struct 1																														
1.1	6	5	7	6	5	7	5	6	6	6	6	6	1	6	6	6	6	6	6	6	5	5	6	5	5.6	1.1	77.1	2.2	100.0	
1.2	1	1	2	3	1	2	1	3	1	2	1	2	1	1	2	1	2	1	2	2	2	1	3	1	1.6	1.0	10.4	0.0	100.0	
1.3	3	5	7	4	1	5	4	6	3	4	4	5	2	1	5	6	5	4	6	5	5	5	5	1	4.2	1.6	53.5	10.7	99.00	
1.4	3	5	7	3	1	4	5	6	3	4	4	5	1	1	5	2	5	4	6	5	5	4	5	2	4.0	1.7	49.3	10.9	99.00	
1.5	1	2	3	2	1	1	1	3	1	2	1	2	1	1	3	3	2	1	3	3	3	2	3	1	1.9	1.0	15.3	0.0	99.00	
1.6	1	3	4	3	1	2	1	2	1	3	2	4	1	1	4	1	3	1	3	3	4	2	3	1	2.3	1.2	20.8	2.5	99.00	
1.7	5	2	6	5	6	6	6	2	1	5	5	6	1	1	3	3	4	6	7	5	6	1	3	1	4.0	2.1	50.0	17.7	99.00	
1.8	1	2	3	2	1	4	1	2	1	2	2	4	1	1	4	2	3	1	2	3	3	1	3	1	2.1	1.1	18.1	1.0	99.00	
1.9	1	2	3	3	1	2	2	3	1	1	2	4	1	1	3	2	3	1	2	3	4	1	2	1	2.0	1.0	17.4	0.0	99.00	
1.10	1	2	4	3	1	2	2	3	1	1	2	4	1	1	3	1	3	1	2	3	3	1	2	1	2.0	1.0	16.7	0.4	99.00	
Quality	5	2	3	6	4	5	4	7	3	3	3	5	1	6	5	5	4	5	5	3	4	5	6	2	4.2	1.5	53.5	7.9		
Context	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (7)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
Cosine Struct 2																														
2.1	7	7	7	7	7	7	7	7	7	7	7	7	3	7	7	7	7	7	7	7	7	7	7	7	6.8	1.0	97.2	0.0	100.0	R
2.2	7	6	6	6	6	7	7	6	7	6	7	7	7	7	7	7	6	7	7	7	7	7	7	7	6.7	1.0	95.1	0.0	100.0	R
2.3	6	6	6	7	5	7	7	6	6	4	7	6	2	3	6	6	5	6	6	6	6	7	7	5	5.7	1.3	78.5	4.5	100.0	
2.4	7	6	6	6	6	7	7	6	7	6	7	7	7	7	7	7	6	6	7	7	7	6	7	7	6.6	1.0	93.8	0.0	100.0	R
2.5	6	5	5	5	4	6	6	4	4	4	7	6	2	3	6	7	5	5	7	6	6	1	6	7	5.1	1.6	68.8	9.5	99.00	
2.6	5	4	3	5	3	6	5	4	2	1	6	6	1	1	4	4	3	5	6	6	6	2	5	3	4.0	1.7	50.0	12.0	99.00	
2.7	5	4	3	4	3	5	4	4	2	2	6	6	1	1	4	3	4	5	6	6	5	1	5	3	3.8	1.6	47.2	10.1	99.00	
2.8	5	4	2	6	3	6	4	4	1	3	5	6	1	1	4	1	4	4	5	5	5	1	4	3	3.6	1.7	43.8	11.5	99.00	
2.9	5	3	3	3	3	5	3	3	1	2	4	6	1	1	4	2	4	5	6	4	5	1	3	1	3.3	1.6	37.5	9.9	99.00	
2.10	5	2	3	3	3	6	3	3	1	2	3	5	1	1	3	1	3	2	4	4	4	1	2	1	2.8	1.4	29.2	7.0	99.00	
Quality	6	7	6	7	5	6	7	7	7	5	7	7	2	2	6	5	5	6	6	6	6	5	7	4	5.7	1.4	78.5	7.1		

Figure B.1: Tabulated human evaluation data for test structures 1 and 2 using the contextual cosine measure that is part of the Chem-DRSM system.

Context																									Avg		Avg		Metnc	
Cosine	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Score	Std	Score	Std	Score	Exact
Struct 3																									(7)		(%)		(%)	
3.1	6	4	6	6	6	6	5	5	3	2	6	6	4	1	3	6	6	5	4	5	6	5	5	3	4.8	1.5	62.5	7.5	97.00	
3.2	6	5	4	6	5	5	6	4	3	2	6	6	4	1	4	5	6	6	5	5	6	2	6	3	4.6	1.5	60.4	8.3	97.00	
3.3	6	5	6	5	5	6	6	6	1	2	6	7	2	1	6	6	6	5	6	6	6	1	6	2	4.8	2.0	62.5	16.4	97.00	
3.4	3	2	2	2	2	3	3	4	1	1	3	4	1	1	2	4	2	1	3	2	5	1	4	1	2.4	1.2	22.9	3.5	96.00	
3.5	3	2	2	2	2	2	2	3	1	1	3	4	1	1	3	5	2	1	3	2	4	3	4	1	2.4	1.1	22.9	2.2	96.00	
3.6	4	5	5	4	4	5	6	4	1	4	5	5	2	1	3	6	5	4	5	5	5	3	5	2	4.1	1.4	51.4	6.9	96.00	
3.7	5	4	5	4	4	5	2	3	1	3	5	5	2	1	4	5	5	3	5	5	5	1	5	1	3.7	1.6	44.4	9.2	96.00	
3.8	3	4	2	3	1	4	1	1	1	1	3	4	2	1	2	4	4	1	4	3	4	3	3	1	2.5	1.3	25.0	4.2	96.00	
3.9	5	5	5	4	4	5	5	1	1	2	5	6	2	1	4	5	5	5	4	5	6	4	4	2	4.0	1.6	49.3	9.6	96.00	
3.10	4	4	5	4	4	5	6	2	1	2	5	5	4	1	4	5	5	6	4	5	6	4	4	1	4.0	1.5	50.0	8.9	95.00	
Quality	4	2	3	5	4	4	5	6	2	2	5	5	3	4	1	5	4	4	4	4	4	4	5	4	3.9	1.2	47.9	3.2		
Context																									Avg		Avg		Metnc	
Cosine	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Score	Std	Score	Std	Score	Exact
Struct 4																									(7)		(%)		(%)	
4.1	1	1	3	4	1	1	1	4	1	2	1	1	2	1	3	3	1	1	1	4	4	1	3	1	1.9	1.2	15.3	3.5	100.0	
4.2	1	1	3	4	1	1	1	4	1	2	1	1	1	1	2	4	1	1	1	3	3	1	3	1	1.8	1.1	13.2	2.4	100.0	
4.3	1	1	3	4	1	1	1	3	1	2	1	1	1	1	2	3	1	1	1	3	4	1	2	1	1.7	1.0	11.8	0.7	100.0	
4.4	1	1	3	4	1	1	1	3	1	2	1	1	1	1	2	3	1	1	1	3	4	1	2	1	1.7	1.0	11.8	0.7	100.0	
4.5	1	1	3	4	1	1	1	3	1	2	1	1	1	1	2	2	1	1	1	3	3	1	2	1	1.6	1.0	10.4	0.0	100.0	
4.6	1	1	3	4	1	1	1	3	1	2	1	1	1	1	2	3	1	1	1	3	3	1	3	1	1.7	1.0	11.8	0.0	100.0	
4.7	1	1	3	4	1	1	1	2	1	2	1	1	1	1	2	2	1	1	1	3	3	1	3	1	1.6	1.0	10.4	0.0	100.0	
4.8	1	2	3	2	1	1	1	3	1	1	1	1	1	1	1	3	2	1	2	5	3	1	2	1	1.7	1.0	11.8	0.7	100.0	
4.9	1	1	4	4	1	1	1	1	1	2	1	1	1	1	2	1	1	1	1	3	5	1	2	1	1.6	1.2	10.4	2.9	100.0	
4.10	1	2	3	3	2	1	1	2	1	1	1	1	1	1	2	2	1	1	2	3	4	1	2	1	1.7	1.0	11.1	0.0	100.0	
Quality	1	4	2	7	2	6	6	4	1	4	2	6	1	5	4	5	5	7	3	4	2	7	6	2	4.0	2.0	50.0	17.0		

Figure B.2: Tabulated human evaluation data for test structures 3 and 4 using the contextual cosine measure that is part of the Chem-DRSM system.

Context																									Avg User Score (7)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
Cosine Struct 5	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421						
5.1	1	1	4	3	1	1	1	3	1	1	2	2	1	1	1	1	3	1	1	5	4	1	2	1	18	12	13.2	3.6	100.0	
5.2	7	7	7	7	7	7	7	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7.0	1.0	99.3	0.0	100.0	R
5.3	7	7	7	7	7	7	7	6	6	7	7	7	7	7	7	7	7	7	7	7	6	7	7	7	6.9	1.0	97.9	0.0	100.0	R
5.4	1	1	3	3	1	1	1	3	1	1	2	2	1	1	1	1	3	1	2	5	4	2	2	1	18	11	13.9	2.2	100.0	
5.5	1	1	2	2	1	3	1	3	1	2	1	2	1	1	2	1	2	1	1	3	3	1	3	1	17	1.0	11.1	0.0	99.00	
5.6	1	1	2	2	1	3	1	3	1	2	1	2	1	1	2	1	2	1	1	3	3	1	3	1	17	1.0	11.1	0.0	99.00	
5.7	1	1	3	2	1	1	1	1	1	1	1	2	1	1	1	1	3	1	2	2	3	1	2	1	15	1.0	7.6	0.0	99.00	
5.8	1	1	3	2	1	1	1	1	1	1	1	2	1	1	1	1	3	1	2	2	3	4	2	1	16	1.0	9.7	0.0	99.00	
5.9	4	4	5	6	2	3	1	2	1	3	1	6	1	1	2	5	3	4	6	6	6	1	4	1	3.3	1.9	37.5	15.7	99.00	
5.10	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	3	1	1	2	2	1	1	1	13	1.0	4.2	0.0	99.00	
Quality	5	5	4	1	3	3	6	7	5	3	4	2	3	1	3	2	3	5	3	3	5	5	5	2	3.7	1.6	44.4	9.2		

Figure B.3: Tabulated human evaluation data for test structure 5 using the contextual cosine measure that is part of the Chem-DRSM system.

Standard Cosine Struct. 1	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (7)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
1.1	6	5	7	7	6	6	6	6	5	6	6	6	2	3	7	6	6	6	6	7	6	7	6	4	5.8	1.2	79.2	3.7	100.0	
1.2	3	5	7	5	5	4	4	6	2	4	5	6	1	1	6	6	5	4	6	6	5	6	6	3	4.6	1.6	60.4	10.6	99.0	
1.3	3	5	6	5	5	4	6	4	2	3	5	6	1	1	6	6	5	4	6	6	5	6	5	3	4.5	1.6	58.3	9.8	99.0	
1.4	1	3	4	4	4	2	3	3	1	2	4	2	1	1	5	5	3	1	5	4	5	4	5	1	3.0	1.5	34.0	8.6	99.0	
1.5	1	3	3	6	3	1	2	2	1	1	1	2	1	1	4	4	3	1	3	3	4	3	4	1	2.4	1.4	23.6	6.3	99.0	
1.6	2	4	5	5	5	7	5	2	3	4	5	3	1	1	6	3	4	2	5	4	6	4	4	5	4.0	1.6	49.3	9.6	99.0	
1.7	2	3	5	5	5	7	5	4	3	4	5	3	1	1	6	3	4	5	5	4	6	4	3	4	4.0	1.5	50.7	8.1	99.0	
1.8	1	3	4	5	2	1	2	2	1	2	1	3	1	1	5	4	3	1	4	3	4	3	5	1	2.6	1.4	26.4	7.4	99.0	
1.9	1	2	4	6	2	1	1	3	1	3	1	3	1	1	4	2	3	1	3	3	4	3	4	1	2.4	1.4	23.6	6.3	99.0	
1.10	1	1	2	4	1	1	4	1	1	1	1	1	1	1	3	3	2	1	2	2	3	2	2	1	1.6	1.0	10.4	0.0	99.0	
Quality	5	3	5	5	6	4	6	5	3	5	4	6	6	1	6	4	5	5	7	6	6	6	5	5	5.0	1.3	66.0	5.0		
Standard Cosine Struct. 2	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (7)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
2.1	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7.0	1.0	100.0	0.0	100.0	R
2.2	7	6	7	7	6	7	7	6	7	6	7	7	7	7	7	7	6	7	7	7	7	7	7	7	6.8	1.0	96.5	0.0	100.0	R
2.3	6	6	6	6	5	6	6	6	6	5	6	6	1	4	6	6	5	6	6	6	6	6	6	6	5.6	1.1	76.4	1.7	100.0	
2.4	7	6	7	7	6	7	7	6	7	6	7	7	7	7	7	7	6	6	7	7	7	7	7	7	6.8	1.0	95.8	0.0	100.0	R
2.5	6	5	5	5	5	6	6	5	1	3	6	6	2	3	6	6	4	5	6	5	6	5	6	6	5.0	1.4	66.0	6.6	98.0	
2.6	6	4	4	5	4	5	5	4	1	2	6	6	1	3	6	5	3	5	6	4	5	4	6	5	4.4	1.5	56.3	8.3	98.0	
2.7	6	4	5	5	4	5	4	4	1	2	6	6	1	3	6	5	4	5	6	4	5	4	5	5	4.4	1.4	56.3	7.3	98.0	
2.8	6	6	6	5	5	6	5	5	1	3	6	6	2	4	6	5	4	5	7	5	6	5	5	6	5.0	1.4	66.7	6.4	98.0	
2.9	6	5	6	6	5	7	7	4	1	3	6	6	2	4	5	5	4	5	7	6	6	6	6	6	5.2	1.5	69.4	8.7	98.0	
2.10	6	5	6	6	3	6	5	4	1	3	6	5	2	4	6	6	4	6	7	5	5	5	6	6	4.9	1.5	65.3	7.9	98.0	
Quality	7	6	5	5	5	6	6	6	5	6	6	7	2	6	5	6	5	6	7	5	6	5	6	6	5.7	1.0	77.8	0.1		

Figure B.4: Tabulated human evaluation data for test structures 1 and 2 using the standard cosine measure that is part of the Chem-DRSM system.

Standard Cosine Struct 3	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
3.1	6	5	5	5	4	6	6	5	1	4	4	4	1	2	6	5	5	1	6	5	5	4	6	4	44	16	56.3	10.2	96.0	
3.2	5	4	5	5	3	6	5	4	1	4	4	5	1	2	4	5	5	5	5	5	5	4	5	1	41	14	51.4	7.4	95.0	
3.3	6	5	6	6	4	6	6	4	1	3	6	6	1	3	6	6	6	5	5	6	5	5	7	2	48	17	63.9	11.4	95.0	
3.4	5	5	6	5	4	6	6	3	1	2	6	6	1	2	5	5	6	6	6	6	6	5	5	2	46	17	59.7	12.4	95.0	
3.5	4	3	4	4	3	2	1	3	1	2	5	3	1	1	4	4	4	3	5	4	3	3	4	1	30	13	33.3	4.8	94.0	
3.6	5	5	6	6	3	5	6	2	1	3	4	6	1	2	5	6	5	4	6	5	6	4	4	4	43	16	55.6	10.1	93.0	
3.7	4	4	4	4	2	5	3	3	1	2	3	6	1	2	3	4	3	2	4	4	3	3	4	3	32	12	36.8	3.0	93.0	
3.8	5	5	5	6	3	6	6	2	1	2	5	6	1	3	5	5	5	5	6	5	5	4	4	6	44	16	56.9	9.8	93.0	
3.9	4	3	4	4	3	5	2	2	1	2	3	3	1	1	3	4	4	2	5	4	3	3	4	2	30	12	33.3	3.0	93.0	
3.10	4	4	4	4	2	6	3	4	1	1	3	3	1	2	5	5	4	2	4	4	3	3	3	3	33	13	37.5	4.9	93.0	
Quality	5	3	4	6	3	3	5	4	1	5	5	3	6	2	4	4	4	3	6	4	4	4	6	4	41	13	51.4	4.7		
Standard Cosine Struct 4	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
4.1	6	5	6	6	5	5	5	6	6	3	6	3	2	2	5	6	6	4	6	6	5	5	5	2	48	14	63.9	6.7	100.0	
4.2	6	5	6	6	5	3	6	6	6	3	7	3	1	6	6	5	6	4	6	6	6	5	5	2	50	15	66.7	8.9	100.0	
4.3	6	4	7	6	5	7	6	6	6	3	7	5	1	5	6	6	6	4	6	6	6	5	6	3	53	14	72.2	7.2	100.0	
4.4	6	4	6	6	5	4	5	7	6	4	7	4	1	5	6	6	6	4	5	7	5	5	6	3	51	14	68.8	6.5	100.0	
4.5	5	3	4	5	4	3	4	4	1	2	4	3	1	1	5	4	5	1	5	4	3	3	5	1	33	15	38.9	7.7	99.0	
4.6	4	4	5	5	4	3	4	4	1	2	4	3	1	2	5	4	5	1	5	4	4	4	5	1	35	14	41.7	6.9	99.0	
4.7	4	4	4	5	4	3	3	4	1	2	4	3	1	2	4	4	5	1	5	4	3	3	5	1	33	13	38.2	5.6	99.0	
4.8	5	3	4	5	5	4	4	5	1	2	5	3	1	3	5	6	5	1	4	4	3	4	5	1	37	15	44.4	8.7	99.0	
4.9	4	4	4	4	5	3	4	4	1	1	4	3	1	2	5	5	5	1	5	4	3	3	5	1	34	15	39.6	7.8	99.0	
4.10	1	1	2	3	1	1	3	3	1	1	3	2	1	1	4	3	3	1	2	2	2	2	4	1	20	10	16.7	0.4	99.0	
Quality	6	5	5	7	5	3	6	7	6	6	6	5	4	3	5	5	6	7	6	6	5	5	6	4	54	11	72.9	1.6		

Figure B.5: Tabulated human evaluation data for test structure 3 and 4 using the standard cosine measure that is part of the Chem-DRSM system.

Standard Cosine Struct. 5	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (7)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
5.1	7	7	7	7	7	7	7	7	7	7	7	7	7	6	7	7	7	7	7	7	7	7	7	7	7.0	1.0	99.3	0.0	100.0	R
5.2	7	7	7	7	6	7	7	7	7	7	7	7	6	6	7	7	7	7	7	7	7	7	7	7	6.9	1.0	97.9	0.0	100.0	R
5.3	1	1	2	3	1	2	1	2	1	1	1	1	1	1	2	1	4	1	1	3	2	2	2	1	1.6	1.0	9.7	0.0	99.0	
5.4	1	1	2	3	1	2	1	2	1	1	1	1	1	1	2	2	1	4	1	1	3	2	2	1	1.6	1.0	10.4	0.0	98.0	
5.5	6	3	5	6	2	6	5	5	1	3	4	6	1	2	6	6	6	4	6	6	5	4	5	3	4.4	1.7	56.9	11.5	98.0	
5.6	1	1	2	1	1	1	1	1	1	1	1	2	1	1	1	1	4	1	2	4	3	2	2	1	1.5	1.0	9.0	0.0	97.0	
5.7	3	2	3	5	1	1	4	3	1	2	2	2	1	1	5	2	6	1	4	4	4	3	2	1	2.6	1.5	27.1	8.3	97.0	
5.8	1	2	2	3	3	3	1	4	1	2	1	2	1	2	3	1	5	1	2	3	2	2	2	1	2.1	1.1	18.1	1.0	97.0	
5.9	1	1	3	2	1	1	1	1	1	1	1	2	1	1	1	1	4	1	2	4	2	2	2	1	1.6	1.0	9.7	0.0	97.0	
5.10	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1	1	4	1	1	3	2	1	2	1	1.4	1.0	6.3	0.0	97.0	
Quality	6	4	3	5	5	5	6	3	7	5	6	5	6	3	5	5	5	6	4	3	5	5	6	6	5.0	1.1	66.0	1.7		

Figure B.6: Tabulated human evaluation data for test structure 5 using the standard cosine measure that is part of the Chem-DRSM system.

Tanimoto (CATI) Struct 1	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
11	1	3	2	5	1	1	1	3	1	2	1	3	1	1	4	2	2	1	2	3	3	2	3	1	20	11	17.4	2.0	100.0	
12	1	3	3	4	1	1	1	3	1	2	3	2	1	1	3	1	4	1	3	3	4	1	3	1	21	12	18.8	2.6	100.0	
13	1	2	2	3	1	1	1	2	1	1	1	1	1	1	2	2	2	1	1	2	2	1	2	1	15	10	7.6	0.0	100.0	
14	1	3	2	5	1	1	1	3	1	2	1	2	1	1	4	1	2	1	2	4	3	4	2	1	20	12	17.4	3.9	100.0	
15	1	3	3	3	1	1	1	3	1	2	2	1	1	2	3	2	3	1	1	2	3	1	3	1	19	10	14.6	0.0	100.0	
16	1	3	4	5	1	2	1	2	1	2	1	2	1	1	5	1	3	1	2	5	4	2	3	1	23	14	20.8	7.0	100.0	
17	1	1	2	4	1	1	1	2	1	1	1	1	1	1	3	3	1	1	1	3	3	1	2	1	16	10	9.7	0.0	100.0	
18	1	3	2	5	1	1	1	2	1	2	1	1	1	1	4	2	2	1	2	4	3	1	2	3	20	12	16.0	2.7	100.0	
19	1	1	4	6	1	1	1	4	1	2	1	1	1	1	5	3	3	1	3	5	5	3	2	1	24	17	22.9	11.1	100.0	
110	1	2	3	3	1	1	1	3	1	2	2	1	1	1	3	2	3	1	2	3	4	1	3	1	19	10	15.3	0.0	100.0	
Quantity	1	3	2	4	1	5	6	5	1	4	1	4	2	1	3	4	4	7	3	4	2	2	4	3	32	17	36.1	11.0		
Tanimoto (CATI) Struct 2	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
21	1	4	4	6	4	4	5	5	1	2	5	6	1	2	6	4	4	2	4	4	5	3	3	2	36	16	43.8	9.7	100.0	
22	2	5	5	6	4	6	6	5	1	3	5	6	2	3	6	5	3	2	5	4	6	5	3	5	43	15	54.9	9.1	100.0	
23	2	1	3	4	2	2	4	6	1	1	3	3	1	1	4	3	2	1	4	4	4	1	2	1	25	14	25.0	6.9	100.0	
24	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	6	7	7	70	10	99.3	0.0	100.0	R
25	7	6	7	7	6	7	7	6	7	6	7	7	7	7	7	7	6	7	7	7	7	7	7	7	6.8	10	96.5	0.0	100.0	R
26	3	2	3	5	4	1	3	4	1	1	3	4	1	2	5	2	2	1	4	4	3	2	5	3	28	13	30.6	5.7	100.0	
27	6	6	7	6	5	6	6	6	4	5	6	7	2	3	6	6	5	5	6	6	6	5	5	6	55	11	74.3	2.4	100.0	
28	6	6	6	6	5	6	5	6	6	5	6	7	2	4	6	6	6	6	6	6	6	5	6	6	55	10	77.1	0.0	100.0	
29	6	6	5	6	5	5	5	5	4	3	6	7	2	5	6	6	5	5	6	5	6	3	5	6	51	12	68.8	2.6	100.0	
210	6	5	5	6	5	5	5	5	4	3	6	7	2	4	6	6	5	5	6	5	4	4	5	6	50	11	66.7	1.7	100.0	
Quantity	4	2	3	4	5	4	6	5	4	2	5	3	2	5	6	5	3	2	3	3	4	5	3	4	38	12	47.2	4.0		

Figure B.7: Tabulated human evaluation data for test structures 1 and 2 using the Tanimoto measure with CATI descriptors that is part of the Chem-DRSM system.

Tanimoto (CATI) Struct 3	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
3.1	4	5	5	5	3	6	6	4	1	3	6	7	1	2	6	5	6	3	6	6	5	5	5	2	4.5	1.7	57.6	12.0	83.0	
3.2	3	5	5	4	3	5	5	3	1	3	5	6	1	2	4	5	3	4	6	5	4	4	4	1	3.8	1.5	46.5	7.9	82.0	
3.3	5	5	5	4	3	5	6	5	1	2	5	6	1	1	4	5	5	4	6	6	5	4	6	2	4.2	1.7	53.5	11.1	82.0	
3.4	5	6	4	4	4	5	6	3	1	5	5	6	1	1	4	5	5	5	6	5	5	4	5	4	4.3	1.5	55.6	8.2	82.0	
3.5	4	5	5	4	3	5	5	6	1	3	5	3	1	2	5	5	5	5	6	3	5	4	4	4	4.1	1.4	51.4	6.3	82.0	
3.6	2	2	2	2	1	2	2	2	1	1	3	3	1	1	3	4	3	1	3	4	3	2	3	1	2.2	1.0	19.4	0.0	80.0	
3.7	2	2	3	2	1	2	2	2	1	1	3	2	1	1	3	3	3	1	2	4	2	2	3	1	2.0	1.0	17.4	0.0	80.0	
3.8	2	3	4	3	1	3	2	2	1	2	3	3	1	2	5	3	2	1	4	3	3	3	3	1	2.5	1.1	25.0	1.1	78.0	
3.9	2	4	2	4	1	4	1	2	1	2	3	3	1	2	6	3	3	1	5	4	3	3	4	1	2.7	1.4	28.5	6.6	78.0	
3.10	6	5	5	5	3	6	6	5	1	5	5	6	1	2	5	6	5	2	5	5	5	4	4	3	4.4	1.6	56.3	9.3	78.0	
Quality	2	4	3	5	4	6	5	3	1	3	6	5	2	2	3	5	5	4	4	4	4	4	5	3	3.8	1.3	47.2	5.1		
Tanimoto (CATI) Struct 4	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
4.1	1	2	4	3	1	2	2	4	1	2	1	2	1	1	5	1	2	1	1	2	4	2	3	1	2.0	1.2	17.4	3.3	100.0	
4.2	1	2	3	3	3	2	1	2	1	1	2	1	1	2	3	2	3	1	1	2	5	2	3	1	2.0	1.0	16.7	0.4	100.0	
4.3	1	2	4	4	3	1	3	2	2	3	3	2	1	2	4	3	4	1	2	2	4	3	3	1	2.5	1.1	25.0	1.1	100.0	
4.4	1	3	4	4	3	1	2	2	3	3	3	2	1	2	5	4	4	1	2	2	5	3	3	1	2.7	1.2	27.8	4.0	100.0	
4.5	1	1	4	3	2	1	2	3	1	2	2	2	1	1	5	2	3	1	1	3	4	2	3	1	2.1	1.2	18.8	2.6	100.0	
4.6	1	3	4	4	2	1	2	4	1	2	1	2	1	1	6	4	4	1	2	2	5	3	4	1	2.5	1.5	25.7	8.4	100.0	
4.7	1	3	3	4	3	2	1	2	3	3	3	1	1	1	4	3	5	1	1	2	4	2	3	1	2.4	1.2	22.9	3.5	100.0	
4.8	2	1	4	3	1	6	1	2	1	2	1	1	1	2	3	2	2	2	1	2	3	2	2	1	2.0	1.2	16.7	3.0	100.0	
4.9	1	5	5	4	3	5	2	3	1	2	3	2	1	2	6	3	6	1	5	5	4	3	3	1	3.2	1.7	36.1	11.0	100.0	
4.10	2	4	3	4	2	5	2	3	3	4	3	2	1	1	6	3	6	1	4	4	4	3	3	3	3.2	1.4	36.1	6.2	100.0	
Quality	1	2	2	3	4	2	4	2	3	5	4	3	3	2	4	4	4	6	2	3	4	3	5	3	3.3	1.2	37.5	3.1		

Figure B.8: Tabulated human evaluation data for test structures 3 and 4 using the Tanimoto measure with CATI descriptors that is part of the Chem-DRSM system.

Tanimoto																									Avg		Avg		Mean	Exact
CAT1	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4231	2413	3124	3214	4312	3421	User	Std	User	Std	Score		
Struct 5	5																							Score	Dev	Score	Dev	(%)		
5.1	7	7	7	7	7	7	7	7	7	7	7	7	7	5	7	7	7	7	7	6	7	7	6	6.8	1.0	97.2	0.0	100.0	R	
5.2	7	7	7	6	6	7	7	6	7	7	7	7	7	6	7	7	7	7	7	7	7	7	7	6.8	1.0	97.2	0.0	100.0	R	
5.3	5	5	4	4	3	4	4	3	1	4	2	2	1	1	5	4	2	2	5	5	2	3	5	1	3.1	1.4	35.4	7.1	100.0	
5.4	5	3	5	4	3	4	4	3	1	4	2	3	1	1	6	5	3	3	5	5	3	3	5	2	3.5	1.4	41.0	6.9	100.0	
5.5	5	2	4	4	1	2	3	3	1	4	2	2	1	1	4	1	4	1	4	4	2	3	4	1	2.6	1.3	27.1	5.8	100.0	
5.6	5	3	6	4	2	5	5	2	1	5	4	3	1	2	6	5	5	4	6	6	5	4	5	2	4.0	1.6	50.0	10.3	100.0	
5.7	4	2	3	4	1	2	2	2	1	3	2	3	1	1	4	1	4	1	4	4	2	2	3	1	2.4	1.2	22.9	2.9	100.0	
5.8	1	1	2	3	1	1	1	1	1	2	1	2	1	1	2	1	2	1	1	3	1	1	2	1	1.4	1.0	6.9	0.0	100.0	
5.9	1	1	2	3	1	1	1	1	1	1	1	2	1	1	2	1	2	1	1	2	2	1	1	1	1.3	1.0	5.6	0.0	83.0	
5.10	1	1	3	3	1	2	1	1	1	2	1	2	1	1	2	1	2	1	3	3	2	2	1	1	1.6	1.0	10.4	0.0	83.0	
Quality	7	4	5	7	5	6	6	5	7	6	6	6	6	4	5	4	6	3	6	6	5	5	6	5	5.5	1.0	74.3	0.3		

Figure B.9: Tabulated human evaluation data for test structure 5 using the Tanimoto measure with CATI descriptors that is part of the Chem-DRSM system.

Tanimoto (FPT)	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
Struct 1																														
1.1	1	3	3	5	1	2	1	3	1	1	3	2	2	1	5	5	6	1	3	4	4	1	4	1	26	16	27.1	10.2	75.0	
1.2	1	1	3	6	1	1	1	2	1	2	1	2	1	1	4	3	2	1	1	4	2	1	3	1	19	13	15.3	5.3	75.0	
1.3	1	2	3	5	1	2	1	2	1	1	1	2	1	1	3	3	3	1	2	3	2	1	2	1	19	10	14.6	0.6	75.0	
1.4	1	2	3	5	1	2	1	2	1	1	1	2	2	1	4	3	2	1	2	3	3	1	3	1	20	11	16.7	1.7	75.0	
1.5	1	1	2	6	1	1	1	2	1	2	1	2	1	1	4	2	1	1	1	3	3	1	3	1	18	13	13.2	4.2	72.0	
1.6	1	1	3	5	1	1	1	2	1	1	1	1	1	1	2	1	2	1	1	3	2	1	3	1	16	10	9.7	0.3	72.0	
1.7	1	1	2	3	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	3	2	1	2	1	14	10	6.3	0.0	72.0	
1.8	1	1	2	5	1	1	1	2	1	1	1	1	1	1	2	1	3	1	1	3	2	1	3	1	16	10	9.7	0.3	72.0	
1.9	1	1	4	4	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	2	1	2	1	14	10	6.9	0.0	72.0	
1.10	1	1	3	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	12	10	3.5	0.0	72.0	
Quality	1	5	2	4	1	6	5	5	1	4	6	6	1	5	3	6	3	7	4	6	6	2	6	5	42	19	52.8	15.8		
Tanimoto (FPT)	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (%)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
Struct 2																														
2.1	4	4	4	5	5	5	4	5	4	2	5	6	2	2	5	3	6	4	7	3	5	3	5	3	43	14	54.2	6.0	100.0	
2.2	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	70	10	100.0	0.0	100.0	R
2.3	7	6	7	7	6	7	7	6	6	6	7	7	7	7	7	6	7	6	7	7	7	7	7	7	67	10	95.1	0.0	100.0	R
2.4	6	5	5	6	5	6	6	5	4	4	6	7	4	2	6	5	5	6	7	6	6	4	6	6	53	11	72.2	2.2	100.0	
2.5	6	5	6	6	6	6	5	5	6	5	6	6	4	2	7	5	5	6	6	6	6	5	6	6	55	10	75.0	0.0	100.0	
2.6	6	4	5	6	4	5	4	4	4	3	6	6	4	2	6	4	6	5	7	5	5	4	5	5	48	11	63.2	2.4	100.0	
2.7	6	4	5	6	5	5	4	4	4	3	6	6	3	2	6	4	5	5	6	5	5	3	5	5	47	11	61.1	2.2	100.0	
2.8	6	4	6	6	6	6	5	5	4	4	6	6	3	2	6	4	6	5	7	6	6	4	5	6	52	12	69.4	3.4	100.0	
2.9	7	6	7	7	6	7	7	6	6	6	7	7	7	7	7	6	7	6	7	7	7	7	7	7	67	10	95.1	0.0	100.0	R
2.10	5	3	5	4	4	4	4	2	4	2	6	6	4	1	5	4	5	4	7	4	4	3	5	3	41	13	51.4	5.8	100.0	
Quality	6	4	4	4	6	3	6	5	5	3	5	5	6	2	5	4	2	5	7	1	6	5	6	5	46	15	59.7	8.4		

Figure B.10: Tabulated human evaluation data for test structures 1 and 2 using the Tanimoto measure with chemical fingerprints that is part of the OpenBabel system.

Tanimoto (FPT) Struct 3	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (\bar{x})	Std Dev	Avg User Score (\bar{x})	Std Dev	Metric Score (%)	Exact
3.1	1	2	3	4	2	2	3	3	1	2	2	2	2	1	5	4	1	1	2	3	3	1	5	1	2.3	1.2	22.2	4.0	92.0	
3.2	1	2	4	4	2	2	2	3	1	2	2	2	2	1	5	4	3	1	3	4	3	2	5	1	2.5	1.3	25.7	4.2	92.0	
3.3	1	2	5	4	3	1	3	2	1	2	2	2	2	1	5	4	3	1	2	4	3	1	5	1	2.5	1.4	25.0	6.4	92.0	
3.4	1	1	2	1	1	1	2	1	1	1	1	1	1	1	2	2	1	1	1	2	2	1	3	1	1.3	1.0	5.6	0.0	91.0	
3.5	1	1	3	2	1	1	2	3	1	1	1	2	2	1	2	2	2	1	1	3	2	2	3	1	1.7	1.0	11.8	0.0	91.0	
3.6	1	1	3	3	1	4	3	1	1	1	1	2	2	1	2	2	2	4	4	4	2	1	2	1	2.0	1.1	17.4	2.0	91.0	
3.7	4	4	5	4	3	4	5	2	1	3	4	3	1	1	6	5	2	2	1	5	5	2	4	2	3.3	1.5	37.5	9.0	91.0	
3.8	1	3	4	4	2	1	3	2	1	2	3	2	1	1	4	4	2	1	2	4	3	1	3	1	2.3	1.2	21.5	2.7	91.0	
3.9	1	1	3	1	1	1	2	2	1	1	1	1	2	1	2	7	1	2	2	2	2	1	2	1	1.7	1.3	11.8	4.5	91.0	
3.10	1	1	3	2	2	1	2	1	1	1	1	1	2	1	3	1	1	1	2	1	1	1	2	1	1.4	1.0	6.9	0.0	91.0	
Quality	2	3	2	4	3	5	5	6	1	4	2	6	2	3	3	4	3	6	3	4	2	6	6	3	3.7	1.6	44.4	9.2		
Tanimoto (FPT) Struct 4	1234	4321	1243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (\bar{x})	Std Dev	Avg User Score (\bar{x})	Std Dev	Metric Score (%)	Exact
4.1	1	3	3	4	6	3	1	4	1	2	1	4	2	2	4	1	2	1	4	2	4	2	5	1	2.6	1.5	27.1	7.8	100.0	
4.2	1	4	4	4	6	4	1	4	1	3	1	4	2	2	4	2	3	1	5	3	4	2	5	3	3.0	1.5	34.0	7.6	100.0	
4.3	6	5	6	6	5	5	5	6	6	5	6	3	3	1	6	2	6	4	4	6	5	2	6	3	4.7	1.6	61.1	9.2	100.0	
4.4	1	3	5	5	5	4	2	5	1	2	4	3	2	1	5	2	5	3	4	3	3	3	5	2	3.3	1.4	37.5	7.0	100.0	
4.5	2	3	3	4	5	4	1	5	1	2	3	4	2	1	5	2	4	1	4	4	3	1	6	2	3.0	1.5	33.3	8.4	100.0	
4.6	2	3	4	5	5	5	1	3	1	1	4	3	2	1	5	2	5	1	5	4	4	2	6	1	3.1	1.7	35.4	11.3	100.0	
4.7	2	3	4	5	5	4	1	3	1	3	4	3	2	1	5	2	5	1	4	4	4	1	6	2	3.1	1.5	35.4	9.0	100.0	
4.8	6	5	5	6	4	6	5	6	6	4	6	3	6	1	6	5	6	2	5	6	5	4	6	2	4.8	1.5	63.9	8.2	100.0	
4.9	2	3	3	5	4	5	2	4	5	3	6	3	4	1	5	4	5	1	3	5	4	2	6	2	3.6	1.5	43.8	7.8	100.0	
4.10	1	2	4	3	2	3	1	4	1	2	2	2	2	1	4	3	4	1	2	3	3	1	4	1	2.3	1.1	22.2	2.2	100.0	
Quality	3	4	3	4	5	3	5	7	3	2	2	6	3	6	4	3	4	6	6	3	4	4	6	4	4.2	1.4	52.8	6.7		

Figure B.11: Tabulated human evaluation data for test structures 3 and 4 using the Tanimoto measure with chemical fingerprints that is part of the OpenBabel system.

Tanimoto (FPT)	1234	4321	*243	1324	1342	2431	1432	2341	2314	2143	3142	3412	1423	2134	3241	4132	4213	4123	4231	2413	3124	3214	4312	3421	Avg User Score (7)	Std Dev	Avg User Score (%)	Std Dev	Metric Score (%)	Exact
Struct 5																														
5.1	7	7	7	7	7	7	7	7	7	7	7	7	5	7	7	7	7	7	7	7	7	7	7	7	6.9	1.0	98.6	0.0	100.0	R
5.2	1	2	3	2	1	2	1	2	1	1	1	2	1	1	2	1	1	1	1	5	2	1	1	1	15	1.0	9.0	0.0	100.0	
5.3	1	2	2	2	1	1	1	2	1	1	1	2	1	1	1	3	1	1	1	4	2	1	1	1	15	1.0	7.6	0.0	100.0	
5.4	7	7	7	7	7	7	7	6	6	7	7	7	7	7	7	7	7	7	7	7	7	6	6	7	6.8	1.0	97.2	0.0	100.0	
5.5	1	2	4	2	1	1	1	2	1	1	1	2	1	1	2	1	1	1	1	3	2	1	1	1	15	1.0	7.6	0.0	100.0	
5.6	4	3	5	5	1	5	5	1	1	3	1	2	1	1	5	4	2	2	3	4	4	2	4	1	2.9	1.6	31.3	9.5	100.0	
5.7	3	2	4	4	1	4	4	3	1	2	1	2	1	1	5	1	4	1	4	4	3	1	3	2	2.5	1.4	25.7	5.8	100.0	
5.8	1	2	5	2	1	1	2	2	1	2	1	1	1	1	2	1	1	1	2	4	2	1	1	1	1.6	1.0	10.4	0.2	100.0	
5.9	1	2	5	2	1	1	2	2	1	2	1	1	1	1	2	2	1	2	2	5	2	1	2	1	1.8	1.1	13.2	1.7	100.0	
5.10	4	3	5	5	1	5	5	2	1	3	1	2	1	1	6	3	2	2	5	5	4	2	4	1	3.0	1.7	34.0	11.3	100.0	
Quality	5	4	4	3	5	2	5	4	5	4	2	3	4	3	2	2	1	2	3	2	3	6	2	5	3.4	1.3	39.6	5.8		

Figure B.12: Tabulated human evaluation data for test structure 5 using the Tanimoto measure with chemical fingerprints that is part of the OpenBabel system.

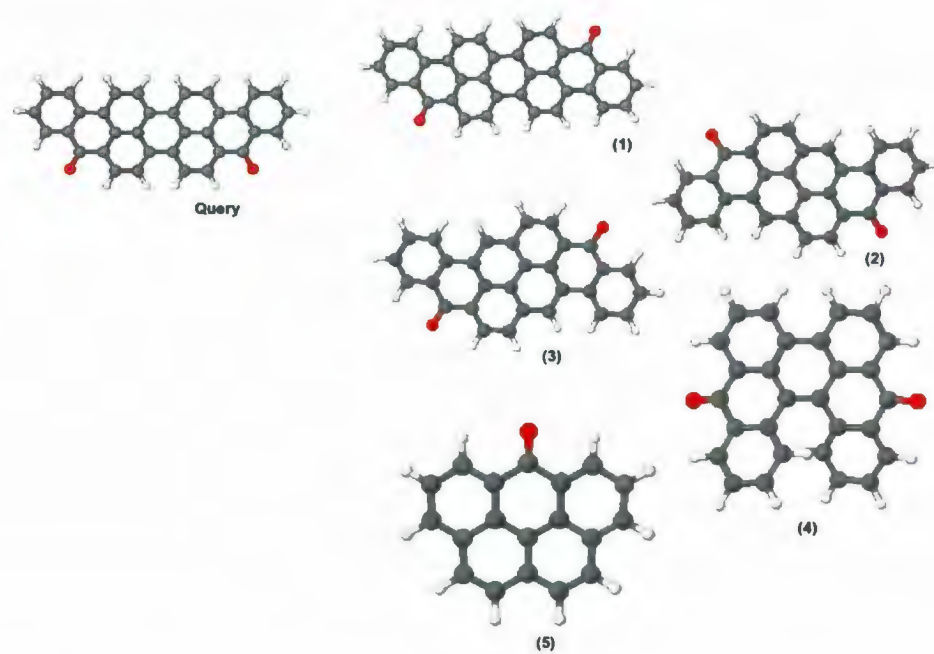


Figure B.13: First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 1 from the human evaluation.

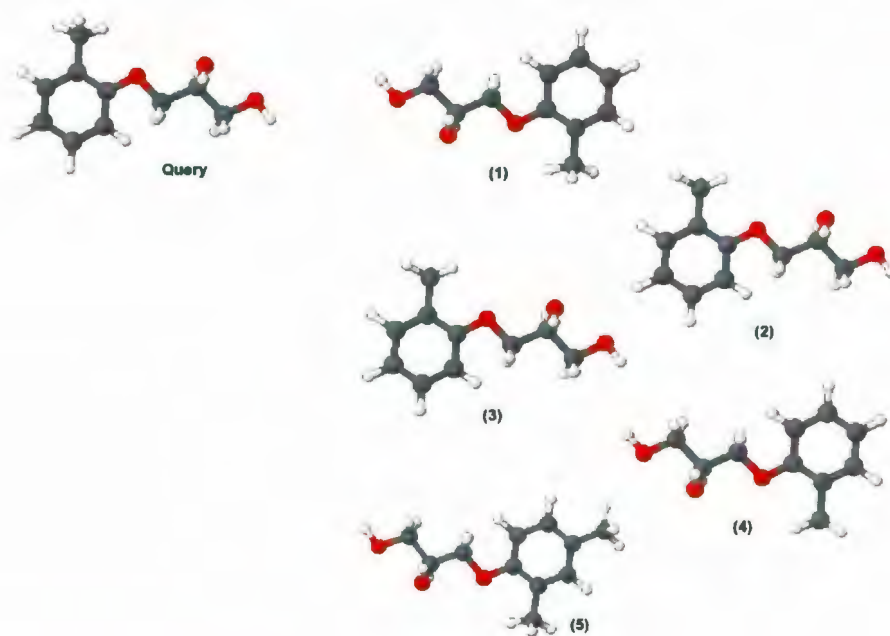


Figure B.14: First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 2 from the human evaluation.

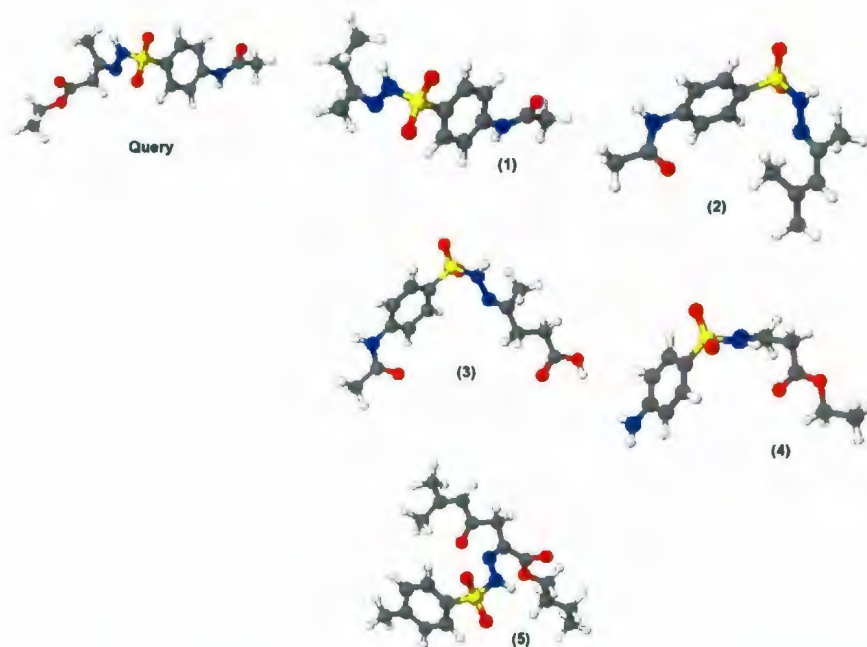


Figure B.15: First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 3 from the human evaluation.

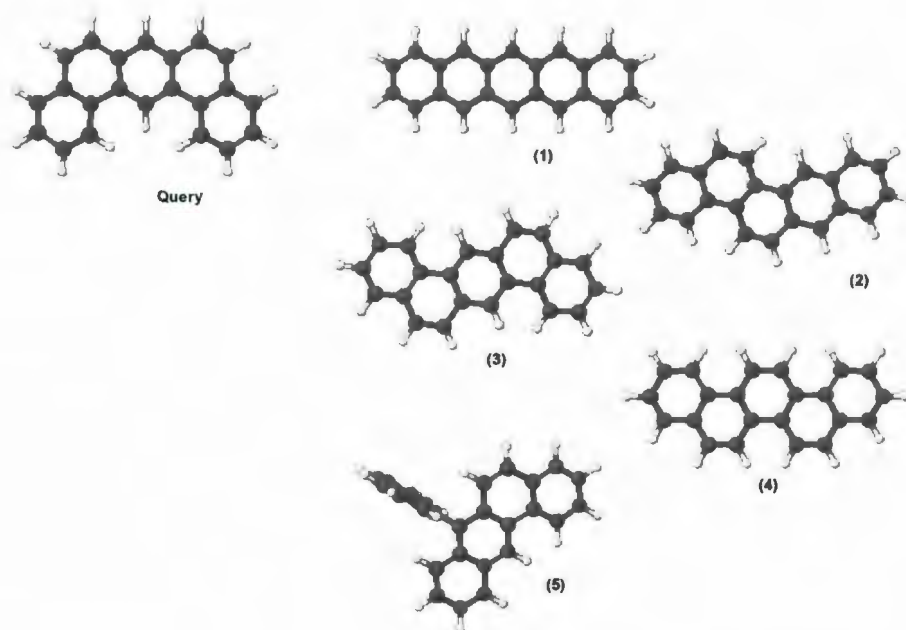


Figure B.16: First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 4 from the human evaluation.

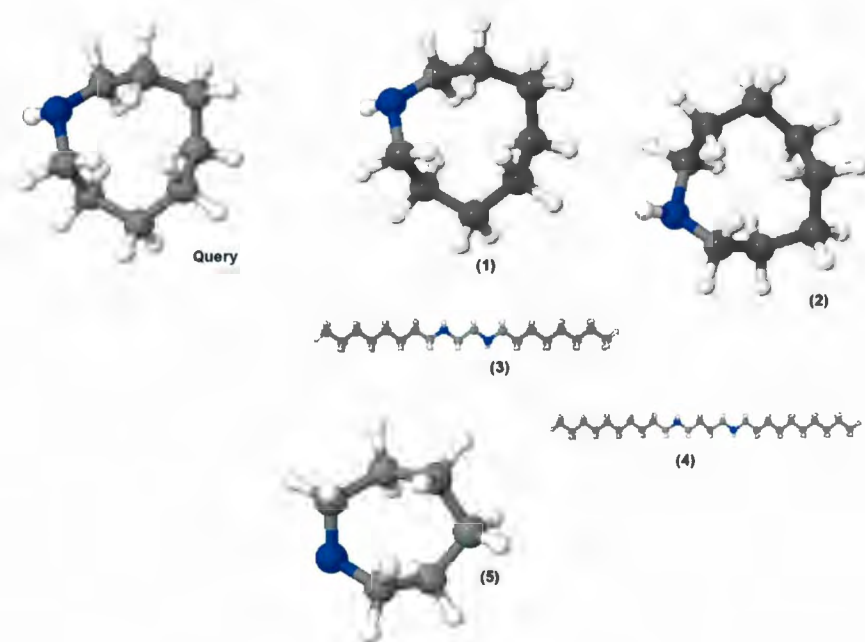


Figure B.17: First five structures returned when using the standard cosine measure (Chem-DRSM system) with query structure 5 from the human evaluation.

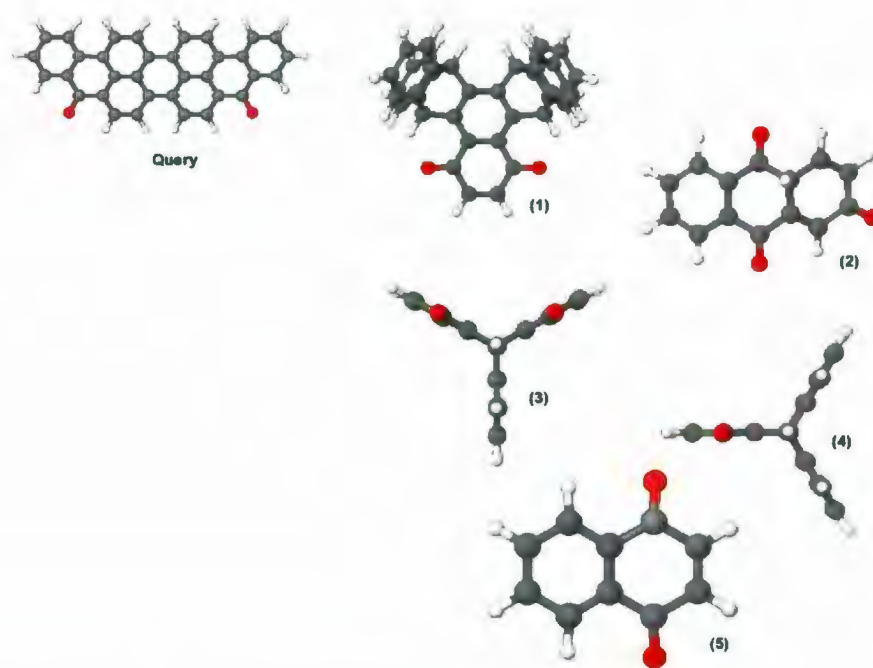


Figure B.18: First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 1 from the human evaluation.

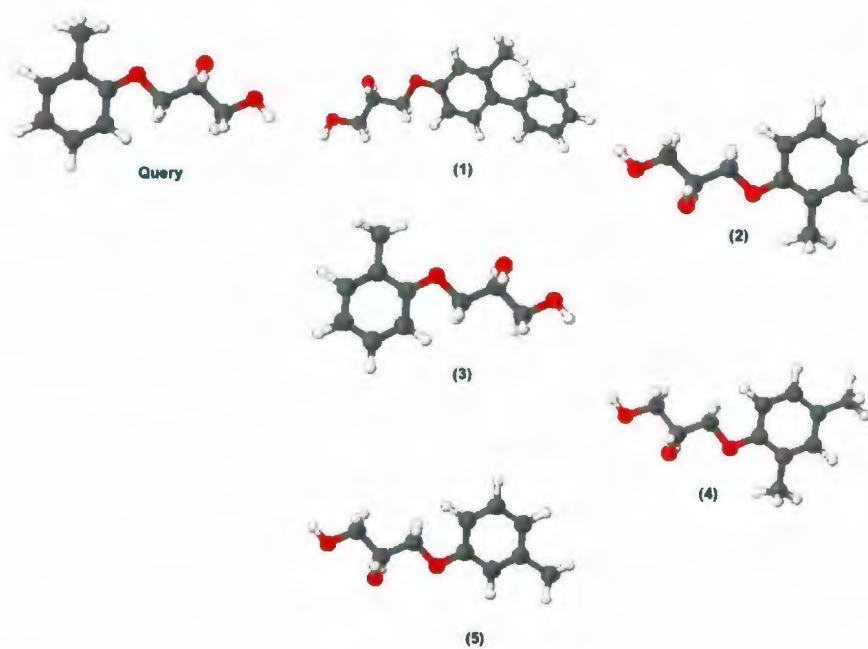


Figure B.19: First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 2 from the human evaluation.

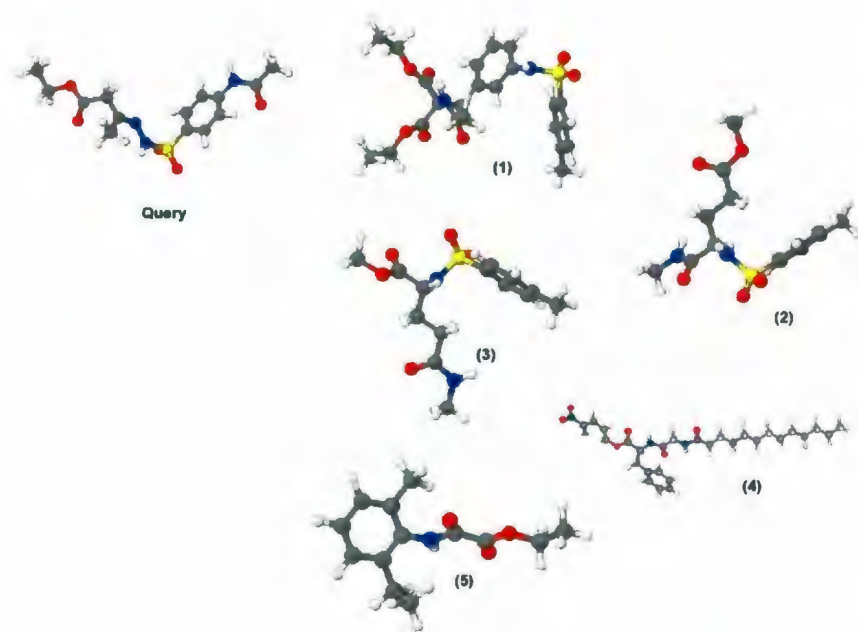


Figure B.20: First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 3 from the human evaluation.

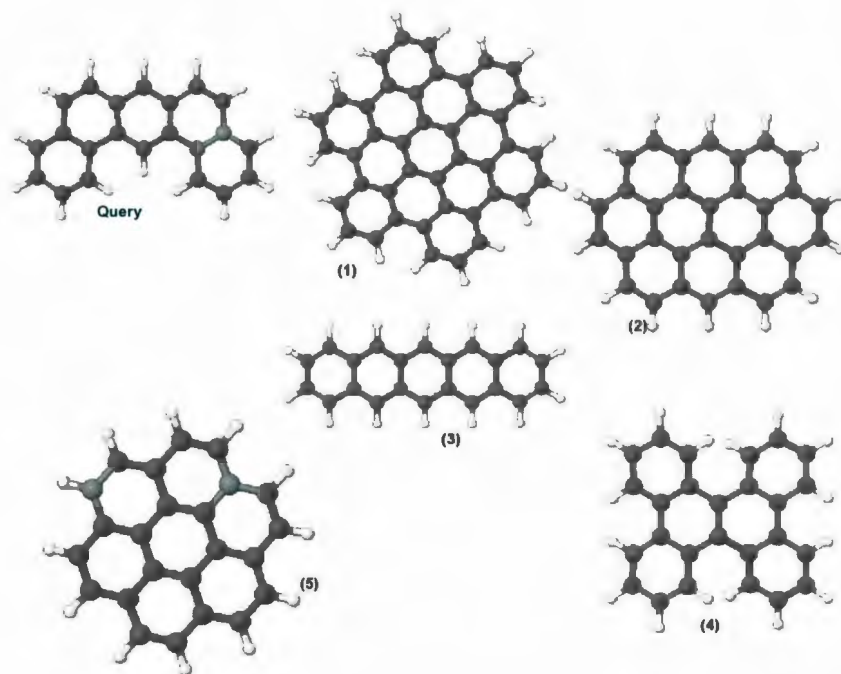


Figure B.21: First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 4 from the human evaluation.

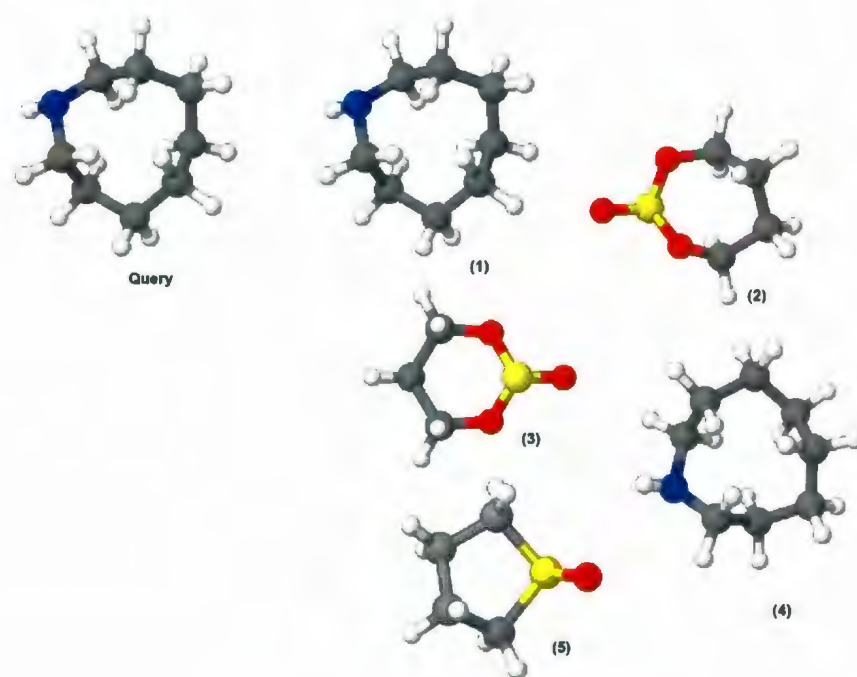


Figure B.22: First five structures returned when using the Tanimoto measure which uses chemical fingerprints (OpenBabel) with query structure 5 from the human evaluation.